VISUAL QUESTION ANSWERING FOR MEDICAL IMAGES WITH EXPLAINABLE AI

A PROJECT REPORT

Submitted By

DEEPANANTH K 195001027

JAYAKRISHNAN S V 195001040

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Department of Computer Science and Engineering

Sri Sivasubramaniya Nadar College of Engineering
(An Autonomous Institution, Affiliated to Anna University)
Kalavakkam - 603110

May 2023

Sri Sivasubramaniya Nadar College of Engineering

(An Autonomous Institution, Affiliated to Anna University)

BONAFIDE CERTIFICATE

Certified that this project report titled **"VISUAL QUESTION ANSWERING FOR MEDICAL IMAGES WITH EXPLAINABLE AI"** is the *bonafide* work of "**DEEPANANTH K (195001027)**, and **JAYAKRISHNAN S V (195001040)**" who carried out the project work under my supervision.
Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

DR. T.T. MIRNALINEE
HEAD OF THE DEPARTMENT
Professor,
Department of CSE,
SSN College of Engineering,
Kalavakkam - 603 110

DR. S. KAVITHA
SUPERVISOR
Associate Professor,
Department of CSE,
SSN College of Engineering,
Kalavakkam - 603 110

Place:
Date:

Submitted for the examination held on. . . . . .

Internal Examiner **External Examiner**

ACKNOWLEDGEMENTS

We thank GOD, the almighty for giving us strength and knowledge to do this project.

We would like to thank and deep sense of gratitude to our guide **DR. S. KAVITHA**, Associate Professor, Department of Computer Science and Engineering, for her valuable advice and suggestions as well as her continued guidance, patience and support that helped us to shape and refine our work.

Our sincere thanks to **DR. T.T. MIRNALINEE**, Professor and Head of the Department of Computer Science and Engineering, for her words of advice and encouragement and we would like to thank our project Coordinator **DR. B. BHARATHI**, Associate Professor, Department of Computer Science and Engineering, members of the project review panel **DR. P. MIRUNALINI**, Associate Professor, Department of Computer Science and Engineering, **DR. B. PRABAVATHY**, Associate Professor, Department of Computer Science and Engineering, **DR. K. LEKSHMI**, Associate Professor, Department of Computer Science and Engineering for their valuable suggestions and support throughout this project.

We express our deep respect to the founder **DR. SHIV NADAR**, Chairman, SSN Institutions. We also express our appreciation to our **DR. V. E. ANNAMALAI**, Principal, for all the help he has rendered during this course of study.

We would like to extend our sincere thanks to all the teaching and non-teaching staffs of our department who have contributed directly and indirectly during the course of our project work. Finally, we would like to thank our parents and friends for their patience, cooperation and moral support throughout our life.

DEEPANANTH K **JAYAKRISHNAN S V**

ABSTRACT

Visual Question Answering (VQA) combines the fields of Natural Language Processing and Computer Vision to generate answers for the questions about the given input image. VQA involves fusion of features extracted from both image and corresponding question, and then the fused feature vector is used for training a Neural Network based model to generate answers. The trained model is then used for generating answers for the given input image and question. In this project, ImageCLEF 2019 VQA-Med Dataset is used. The images from this dataset are complex to analyze and are of low resolution. Each image from the dataset has multiple questions. For VQA model creation, VGGNet and Bidirectional Encoder Representations from Transformers (BERT) are used for extracting features from images and questions respectively. These features are concatenated and the answer generation is achieved using BERT. The trained model is validated using the test dataset and it resulted in 46.8% accuracy, 48.61 BLEU Score and 50.97 WBSS Score. In addition to this, the outcome of the VQA model is analyzed using Explainable AI (XAI) techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). Combining XAI with VQA for the medical images gives analysis of the answer and supports the generated answer with justifications.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

VQA	Visual Question Answering
XAI	Explainable Artificial Intelligence
LSTM	Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
LIME	Local Interpretable Model-agnostic Explanations
SHAP	SHapley Additive exPlanations
CIU	Contextual Importance and Utility
Grad-CAM	Gradient-weighted Class Activation Mapping
BLEU	BiLingual Evaluation Understudy
WBSS	Word Based Semantic Similarity

CHAPTER 1

INTRODUCTION

Artificial Intelligence has grown exponentially over the past 10-15 years. The intelligent models or agents have solved many real-world problems and were able to learn or identify patterns among different kinds of data and provide the desired output. Now moving on to the next phase of learning, where the model tries to answer the questions asked by the user related to some data provided along with the question. Visual Question Answering (VQA) is an emerging task in the field of Artificial Intelligence and Computer Vision that aims to generate answers for the given questions by looking into the given image which corresponds to the question. VQA can be applied to various types of images like Natural Images, Medical Images or Cartoon Images. In this project, we aim to use different types of medical images like radiology images, CT scans, MRI scans etc., along with relevant questions and try to generate answers. Features are extracted from both image and question. The features are fused and are used to train a Neural Network architecture. The trained Neural Network architecture is used to generated answer for the input image and question. In order to justify the answer, some explanations are needed. There should be some features that correspond to the answer that is generated. Such features can be analyzed and identified using Explainable AI (XAI) tools for providing explanations.

1.1 MOTIVATION

The medical domain is one, where there are new viruses and new diseases that emerge and also the one that needs faster analysis of patients' conditions. Applying Artificial Intelligence techniques in the Medical field is more effective, where deep analysis of problems can be performed with the help of different AI techniques or algorithms. VQA in the medical domain would help doctors to analyze and get in-depth knowledge of medical images. The doctors can also submit their queries and get the required information [26]. Not only doctors, but even patients could use these VQA tools to get answers to their questions. Instead of searching and reading unknown articles from various websites, they can use these tools to get required information. There are Visual Question Answering models [1–4, 7, 9–11, 16, 19, 20, 24] for the medical domain that could generate answers for questions, but they do not give any justification for the predicted outcome. To overcome this limitation, the proposed model uses a XAI technique to justify the outcome of the VQA model.

1.2 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 XAI tools like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) to provide explanations on the outcome.

Input : The input is the medical images of different modalities and planes for different organs and their relevant questions.

Output : For the given image and the query the proposed system predicts the answer which will be validated using XAI technique. A sample set of images and queries with the corresponding answers is shown in Figure 1.1.

(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 1.1: Sample Images and Questions with Corresponding Answers from ImageCLEF 2019 VQA-Med Dataset (Image IDs: synpic191614, synpic28495)

Objectives

- To collect and analyze the CLEF2019 VQA dataset.
- To build an efficient VQA model that generates the answers to the questions related to the given Medical Image.
- To analyze the generated answer using XAI tools for providing explanations.

1.3 SYSTEM REQUIREMENTS

Hardware Requirements

- Machines with Intel (i5, i7 or xeon) or AMD (Ryzen 3, Ryzen 5) processors with a minimum of 8GB of RAM and 128GB of storage.
- Nvidia GPU for hardware acceleration

Software Requirements

- Python 3.5 or higher
- Compatible Nvidia GPU Drivers and Cuda Toolkit
- Pytorch, Tensorflow >=2.8
- LIME, SHAP

1.4 ORGANISATION OF REPORT

The report is organized as follows: brief introduction, motivation for the project and the proposed problem statement are discussed in Chapter 1. Some of the existing systems for VQA and XAI are discussed in Chapter 2. In Chapter 3, the design of the proposed system with modules, the algorithms used and implementation are explained. Chapter 4 illustrates the results of the implementation and performance analysis. The conclusion and future work are summarized in Chapter 5.

CHAPTER 2

LITERATURE SURVEY

This section discusses various research papers on Visual Question Answering (VQA) for medical images with its techniques and limitations. This section also discusses about various Explainable AI (XAI) techniques used for Deep Learning algorithms in medical domain.

2.1 VISUAL QUESTION ANSWERING

Visual Question Answering (VQA) task involves generating answers for the given question provided the corresponding input image. VQA can be applied to several types of images like natural images, cartoon images and medical images. Applying VQA for medical images is challenging as medical images in general are complex and are of low resolution.

ImageCLEF forum is an international forum which focuses on conducting tasks involving images, which are to be solved using Machine Learning (ML) / Deep Learning (DL) algorithms. VQA tasks have been posted from 2018 and in this project the dataset of VQA-Med 2019 [6] is used. The dataset has four categories of questions such as Modality, Plane, Organ and Abnormality. For each image in the dataset, there are 4 questions of each category.

Aisha Al-Sadi et al., [3] had used the characteristic of having different categories of questions in the dataset and had built ensembles of Convolution Neural

Network (CNN) models for each category of the questions which had resulted with an accuracy of 60.8% and a BiLingual Evaluation Understudy (BLEU) Score of 63.4.

Lubna A et al., [1] considered only the modality related questions with specific modality categories in the dataset. A CNN model is built and trained for predicting the modality of the images. This resulted with the accuracy of 83.8%.

Instead of building different classification models for different categories of data, Rabia Bounaama et al., [7] proposed an approach in which the features are extracted from both image and question respectively and these features are fed to another model for predicting the answer. The image and question features are extracted using a pre-trained VGGNet and Long Short-Term Memory (LSTM) respectively. These features are concatenated and are fed to a LSTM model to predict the answer by classification approach which resulted with an accuracy of 46.2% and a BLEU Score of 48.6.

An encoder-decoder model is proposed by Imane Allaouzi et al., [11] where the image features are extracted by using DenseNet and are then concatenated with the question features extracted using LSTM. These features are fed to a fully connected neural network to predict the answer. The answer generated in the previous iteration is then concatenated with the existing features and again fed to the same fully connected neural network to predict the next word in the answer recursively. This resulted with 55.6% accuracy and BLEU Score of 58.3.

Aisha Al-Sadi et al., [4] used the idea of encoder-decoder based approach for generating answers for abnormality type questions. For organ and plane related questions, the VGGNet architecture is used to predict the answers by

classification. The modality questions were handled by first predicting the major categories like MRI, CT, XR, Ultrasound, etc., and different models for each of these major categories of modality are used for finding the subcategories. This approach resulted with 57% accuracy and a BLEU Score of 59.1.

Yangyang Zhou et. al., [24] used Inception-ResNet-152 to extract features from the images and Bidirectional Encoder Representations from Transformers (BERT) to extract question features. A Multi-Layer Perceptron is used to unify the dimensions of the features and are then concatenated. For modality, plane and organ type of questions, these features are fed to a classification layer to predict the answer. For abnormality type questions, a generative approach is proposed which involves predicting the next words of the answer recursively. The last generated word is encoded with the features for predicting the next word in the next iteration. This resulted with 60.6% accuracy and 63.3 BLEU Score.

Instead of just concatenating the image and question features, various feature fusing techniques are used for VQA tasks. Dhruv Sharma et al., [9] used Multimodal Factorized Bilinear Pooling (MFB) feature fusion technique to fuse the image feature extracted using ResNet-152 and the question features extracted using pre-trained BERT. Attention mechanisms like Image Attention and Image-Question Co-attention are used in their proposed model. An LSTM is used for answer generation similar to the answer generation technique proposed by Imane Allaouzi et al., [11]. This resulted with an accuracy of 63.8%.

Abhishek Thanki et al., [2] used Element-wise multiplication technique for fusing the image and question features extracted using VGGNet-19 and LSTM respectively. These fused features are then fed to an LSTM to predict the answers

recursively. The approach resulted with an accuracy of 15.5% and a BLEU Score of 45.5.

Lei Shi et al., [20] experimented with feature fusion techniques for the VQA tasks. The image and question features extracted using ResNet-152 and Bi-LSTM are fused with Multi-modal Factorized High-order pooling (MFH). The fused feature vector is then used for predicting the answer. For modality, plane and organ related questions, single label classification models is used to generate answer. For abnormality, multi-label classification model is used where each of the output labels corresponds to the part of the answer. This resulted with 56.6% accuracy and 59.3 BLEU Score.

Attention mechanisms are used with VQA tasks to extract the image feature based on the question. Shengyan Liu et al., [19] used an attention module which takes the image features extracted using pre-trained Xception model and question features extracted using Gated Recurrent Units (GRU) as input and outputs a new set of global image features which is then used along with repeated question vectors. These features are fed to a softmax layer for answer prediction. This approach with attention module resulted with 21% accuracy and a BLEU Score of 39.3.

Minh H. Vu et. al., [16] used an attention module on the image features extracted using ResNet-152 to get the global image features that contribute to the corresponding question. The question features are extracted using a pre-trained BERT. A bi-linear transformation techniques is used on these features. The resultant feature is used to predict the answer. This approach with attention module resulted with 61.60% accuracy and 63.89 BLEU Score.

Transformer models like BERT had been proposed for answer generation by Fuji Ren et al., [10] where different models for different types of questions including different models for closed-ended questions were built. For open-ended questions like finding the modality, plane and organ, a classification BERT model was used. For questions which deal with the abnormality in the image, a generative model is trained for generating the answer by masking random words and predicting them, resulting with the accuracy of 64% and BLEU Score of 65.9.

The previous works of VQA task involving ImageCLEF 2019 dataset is summarized in the Table 2.1.

TABLE 2.1: Literature Survey - VQA

Paper Title	**Methodology**	**Limitations**
Visual question answering in the medical domain based on deep learning approaches: A comprehensive study [3]	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	The type of the questions in the real time is unknown and hence choosing the model for a corresponding question would be another major task.
MoBVQA: A Modality based Medical Image Visual Question Answering System [1]	Only Modality related questions are considered and a CNN model is trained to predict the modality of the image. Dataset: ImageCLEF 2019 VQA-Med Analysis: Accuracy-83.8	Only Modality based questions are considered and within that only major categories of modalities are predicted.

Tlemcen University at ImageCLEF 2019 Visual Question Answering Task [7]	Dataset: ImageCLEF 2019 VQA-Med Image Feature Extraction: VGGNet16 Question Feature Extraction: LSTM Answer Generation: LSTM Analysis: Accuracy-46.2, BLEU-48.6	Transformer based models for question feature extraction and answer generation would perform well compared to LSTM.
An Encoder-Decoder model for visual question answering in the medical domain [11]	Dataset: ImageCLEF 2019 VQA-Med Image Feature Extraction: DenseNet-121 Question Feature Extraction: LSTM Feature Fusion: Feature Concatenation Answer Generation: Fully Connected Neural Network Analysis: Accuracy-55.6, BLEU Score-58.3	For each query, entire image needs to be looked up, while a attention based mechanisms could be used to look up only the question centric regions.

JUST at ImageCLEF 2019 Visual Question Answering in the Medical Domain [4]	Dataset: ImageCLEF 2019 VQA-Med Modality, Plane and Organ categories: VGGNet classification for each categories. Abnormality: Image is fed into LSTM and a set of features from hidden layer is fed to another LSTM then for a softmax layer for prediction. Analysis: Accuracy-57, BLEU-59.1	Abnormality based questions resulted with very low accuracy.
TUA1 at ImageCLEF 2019 VQA-Med: A classification and generation model based on transfer learning [24]	Dataset: ImageCLEF 2019 VQA-Med Image Feature Extraction: Inception-Resnet-v2 Question Feature Extraction: BERT Feature Fusion: MLP for dimensions mapping & concatenating Answer Generation: A neural network for classification Analysis: Accuracy-46.2, BLEU-48.6	Abnormality based questions resulted with very low accuracy.

MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain [9]	Dataset: ImageCLEF 2019 VQA-Med, PathVQA Image Feature Extraction: ResNet152 Question Feature Extraction: BERT Feature Fusion: Multimodal Compact Bilinear Pooling (MCB) Answer Generation: LSTM Analysis: Accuracy-63.6	Abnormality based questions were not considered.
MIT Manipal at ImageCLEF 2019 Visual Question Answering in Medical Domain [2]	Dataset: ImageCLEF 2019 VQA-Med Image Feature Extraction: VGGNet-19 Question Feature Extraction: LSTM Feature Fusion: Element-wise multiplication Answer Generation: LSTM Analysis: Accuracy-15.8, BLEU-45.5	Attention based techniques for feature fusion could be used to increase the accuracy.

Deep Multimodal Learning for Medical Visual Question Answering [20]	Dataset: ImageCLEF 2019 VQA-Med Image Feature Extraction: ResNet-152 Question Feature Extraction: Bi-LSTM Feature Fusion: Multi-modal Factorized High-order pooling(MFH) Answer Generation: A neural network for single label and multi-label classification Analysis: Accuracy-56.6, BLEU-59.3	For answer generation, a transformer based model could be used for efficient output.
An Xception-GRU Model for Visual Question Answering in the Medical Domain [19]	Dataset: ImageCLEF 2019 VQA-Med Image Feature Extraction: Xception Model Question Feature Extraction: Gated Recurrent Unit (GRU) Feature Fusion: Attention module Answer Generation: Softmax layer Analysis: Accuracy-21, BLEU-39.3	Efiicient attention module can be used for extracting best features from the image.

Ensemble of Streamlined Bilinear Visual Question Answering Models for the ImageCLEF 2019 Challenge in the Medical Domain [16]	Dataset: ImageCLEF 2019 VQA-Med Image Feature Extraction: ResNet-152 Question Feature Extraction: Pre-trained BERT Feature Fusion: Attention mechanism (MLB) Answer Generation: Bilinear transformation and softmax layer Analysis: Accuracy-61.60, BLEU-63.89	Abnormality based questions resulted with very low accuracy.
CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering [10]	Dataset: ImageCLEF 2019 VQA-Med Image Feature Extraction: ResNet-152 Question Feature Extraction: BERT Tokenizer Answer Generation: BERT Analysis: Accuracy-64, BLEU Score-65.9	The proposed solution for VQA is building different models for different types of question such as Modality, Plane, Organ and Abnormality. But in reality, the exact type of a question may not be known.

2.2 EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

Explainable AI is a domain that deals with techniques and tools that helps to explain the predictions made by ML/DL models [22]. These tools analyze the model and find the reasons behind the prediction made by the model.

Explainable AI in medical domain helps to achieve:

- **Increased transparency:** XAI techniques gives explanations on why an AI system has arrived at an outcome or prediction. Since it gives such explanations, transparency of the AI systems increases and can increase the trust on that system [22].
- **Result tracing:** XAI techniques helps to trace the features of the input that affect the outcome of the AI system [22].
- **Model improvement:** In general, AI systems learn from some information in the training data. There are chances where the learned rules are erroneous and can lead to incorrect predictions. Hence XAI techniques can be applied to the AI systems and can be validated if the model has learned correct information from the data [22].

There are several XAI techniques that have been widely used for different types of architecture. Knapič S et al., [13] analyzed the predictions made by a CNN model trained to classify the bleeding and non-bleeding endoscopy image of the gastrointestinal tract. XAI techniques like Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), Contextual

Importance and Utility (CIU) are used in the proposed system for generating explanations. The predictions of the model were analyzed and the analysis of the three XAI techniques were compared. The comparison showed that CIU outperformed both LIME and SHAP.

Cameron Severn et. al., [8] applied SHAP for their prediction model for explaining the results. TCGA-GBM Dataset is used which has MR images of adult diffuse gliomas. The radiomic features are extracted from these images. These radiomic features are then used for training XGBoost and LightGBM models. These models are then interpreted using SHAP which gives the analysis of the radiomic features contributing to the outcome.

Avleen Malhi et. al., [5] used the Red Lesion Endoscopy dataset and trained a CNN based model for classifying bleeding and non-bleeding images. LIME is used for explaining the predictions made by the CNN model by marking the bleeding regions of the image.

Apart from the CIU, LIME and SHAP tools, there are several other XAI techniques like Class Activation Mapping (CAM), Gradient-weighted Class Activation Mapping (Grad-CAM), Anchors and Integrated Gradients.

CAM is helpful in explaining the models' predictions by generating a heatmap that represents the contribution of features of the image to the predicted outcome. Jannis Born et. al., [12] used a VGGNet model for the Lung Ultrasound Images (LUS) to detect COVID-19, bacterial pneumonia and non-COVID-19 viral pneumonia. The VGGNet model is then interpreted using CAM to generate heatmaps that justifies the predicted outcome.

Similarly, Yu-Huan Wu et. al., [25] used a ResNet model for classifying COVID-19 CT scan images. The explainabilty is provided using CAM.

Ramprasaath R. Selvaraju et al., [17] proposed a technique called Grad-CAM for explaining and understanding CNN based models. This technique helps to visualize the important regions on the input image, that corresponds to the predicted outcome. This technique is primarily used to understand CNN-based models such as Image-Captioning and VQA models.

Integrated Gradients and Anchors techniques are combined with LIME and SHAP for analyzing and providing explanations for a model that detects COVID-19 from the X-ray images [21]. The results of each of the above tools are combined to give explanations.

The discussion on various XAI techniques used in various DL/ML models for medical domain is summarized in Table 2.2.

TABLE 2.2: Literature Survey - XAI

Paper Title	**Dataset**	**Model**	**XAI technique**
Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain [13]	Red Lesion Endoscopy Dataset	CNN	LIME, SHAP, CIU
A Pipeline for the Implementation and Visualization of Explainable Machine Learning for Medical Imaging Using Radiomics Features [8]	TCGA-GBM Dataset	XGBoost and LightGBM	SHAP
Explaining Machine Learning-Based Classifications of In-Vivo Gastral Images [5]	Red Lesion Endoscopy Dataset	CNN	LIME
Accelerating Detection of Lung Pathologies with Explainable Ultrasound Image Analysis [12]	Lung Ultrasound (LUS)	VGGNet	Class Activation Mapping (CAM)

JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation [25]	COVID-19 Classification and Segmentation (COVID-CS) dataset	ResNet	Class Activation Mapping (CAM)
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization [17]	-	-	Gradient-weighted Class Activation Mapping (Grad-CAM)
LISA : Enhance the explainability of medical images unifying current XAI techniques [21]	COVID-19 Dataset	CNN (Pre-trained)	LIME, SHAP, Anchors, Integrated Gradients

2.3 INFERENCE

Pre-trained models such as VGGNet, ResNet and DenseNet are majorly used for extracting features from images. LSTM and BERT are primarily used for extracting question features. For answer generation, neural network based models like LSTM, a simple Fully Connected Neural Network or BERT are used. The VGGNet and LSTM combination is widely used for VQA tasks. In this project, VGGNet and BERT combination is used, which were not used together for VQA tasks. BERT is a popular NLP model and is used for answer generation as it is efficient for Masked Language Modeling (MLM) tasks.

Though there are existing VQA models [1–4, 7, 9–11, 16, 19, 20, 24] that are developed, these models do not provide any explanations. Applying XAI techniques to the ML/DL models, not only provides analysis on the prediction, but also helps us to improve the developed model. Several XAI techniques like LIME, SHAP, CIU and Grad-CAM are available and out of these LIME and SHAP are used to generate explanations in this project.

CHAPTER 3

PROPOSED SYSTEM

The proposed system aims to develop a Visual Question Answering (VQA) model for the ImageCLEF 2019 VQA-Med Dataset using VGGNet and Bidirectional Encoder Representations from Transformers (BERT). The ImageCLEF 2019 VQA-Med Dataset is chosen for this project, since it has four categories of questions and also it has images with different modalities (like CT, MRI, Ultrasound and etc.,.), different planes (like axial, lateral, sagittal and etc.,.) and different organs (like lung, skull, spine, muskuloskeletal and etc.,.). These images are complex to analyze and are of low resolution. VGGNet and BERT are used in this project for extracting features from images and questions respectively. A BERT model is trained for Masked Language Modeling (MLM) which generates answers for the corresponding input by predicting the masked words. Further, the results of the VQA model are to be analyzed using Explainable AI (XAI) techniques like Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). The detailed system architecture diagram is shown in Figure 3.1.

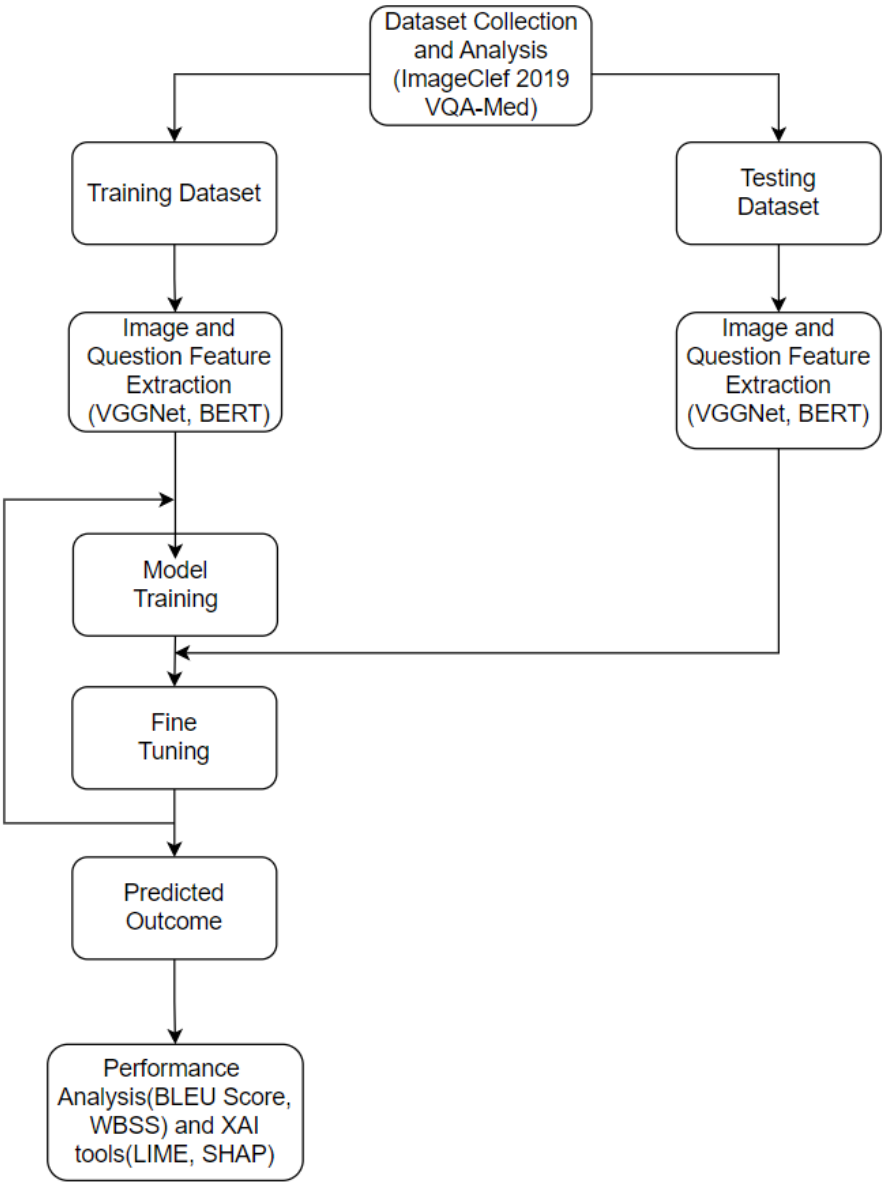


FIGURE 3.1: System Design

3.1 DATASET DESCRIPTION

ImageClef 2019 VQA-Med Dataset is used in this project. The dataset is collected and analyzed in terms of the categories of question such as Modality, Plane, Organ and Abnormality. Within these categories of questions, a pivot table is used to

analyze the classes available under each categories. The analysis of the dataset is further discussed under the Section 4.1.

3.2 FEATURE EXTRACTION

VQA involves extracting features from both image and the question. Before extracting features from the images and questions, it is necessary to pre-process them. Pre-processing of images & questions and also the feature extraction from them are discussed under this section.

3.2.1 Image Pre-Processing

The images are pre-processed by resizing the image to a constant size of (224,224). The resized images are then used for Image Feature Extraction. The function for image pre-processing is summarized in Algorithm 1.

Algorithm 1 Image Pre-Processing

Input : $Image\ of\ different\ size$ ▷ Input Image

Output : $Resized\_Image\ of\ size\ 224 \times 224$ ▷ Resized Image

function IMAGEPREPROCESS(Image)

$Resized\_Image \leftarrow cv2\_resize(Image, (224, 224))$

return *Resized_Image*

end function

3.2.2 Image Feature Extraction

A pre-trained VGGNet is used to extract features from the images. The last layer of the VGGNet model is replaced with a dense layer of 960 units. When an image is given to this VGGNet Model, the values at the newly added dense layer are the required image features.

The architecture of VGGNet which is modified by replacing the last layer is shown in Figure 3.2.

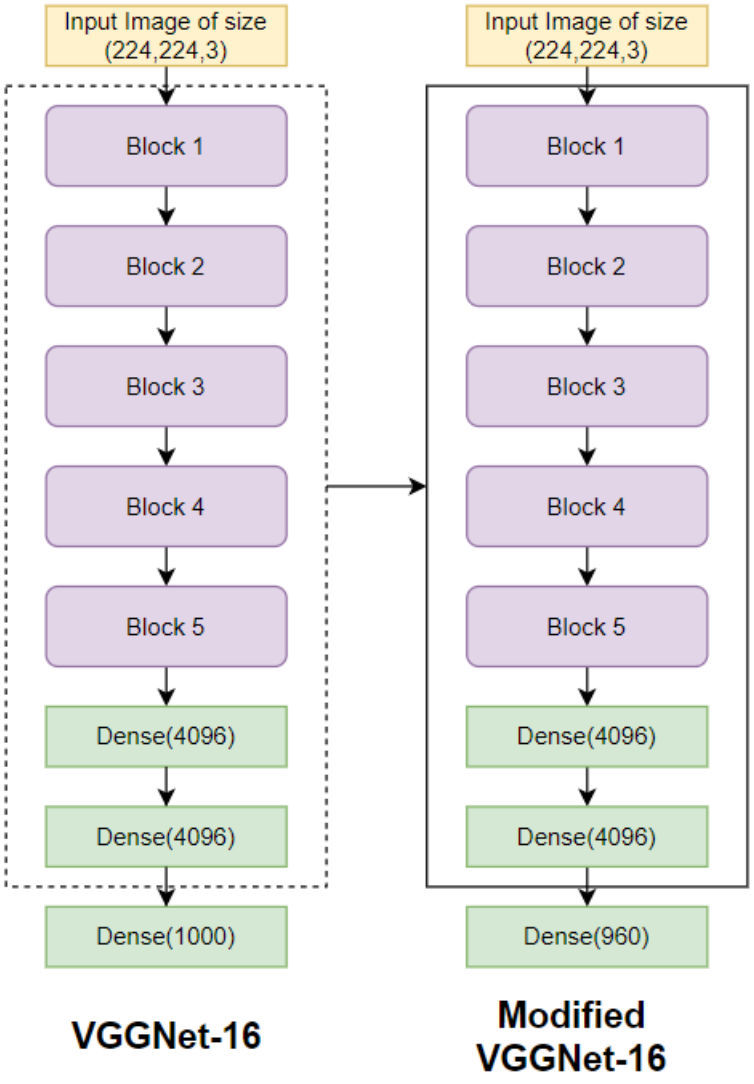


FIGURE 3.2: Modified VGGNet Architecture

The summary of the modified VGGNet model is shown in Figure 3.3.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
new_fc (Dense)	(None, 960)	3933120

FIGURE 3.3: Model Summary of VGGNet

The values from the newly added layer are added with 100 and are rounded to its highest integer value as BERT does not accept float values. The rounded values are the encoded image features. The values are added with 100 so that the encoded

feature values do not match the token Ids of the special tokens. This image feature encoding is depicted in the Algorithm 2.

Algorithm 2 Image Feature Encoding

Input : $Image$ ▷ Input Image

Output : $Image\_Encoding\ of\ length\ 960$ ▷ Encoded Image Features

$VGGModel \leftarrow VGG16()$

$VGGModel.layers[-1] \leftarrow Dense(units = 960, activation = "relu")$

function GETIMAGEENCODING(Image)

$Preprocessed\_Image \leftarrow imagePreProcess(Image)$

$Image\_Features \leftarrow VGGModel(Preprocessed\_Image).layers[-1].values$

$Image\_Encoding \leftarrow list((ceil(x) + 100)\ for\ x\ in\ Image\_Features)$

return $Image\_Encoding$

end function

3.2.3 Question Pre-Processing

The text data usually is associated with punctuation and special characters and hence the question needs to be pre-processed by removing these special characters and punctuation. Also, the text is converted to lower case. The question pre-processing is summarized as Algorithm 3.

Algorithm 3 Question Pre-Processing

Input : *Question* ▷ Input Question for pre-procesing

Output : *Preprocessed_Question* ▷ Preprocessed Question

function TEXTPREPROCESS(Question)

Lower_Case_Qn ← *Question.lower*()

Preprocessed_Question ← *Lower_Case_Qn.replace*("[^ *a* − *z*0 − 9]", " ")

return *Preprocessed_Question*

end function

3.2.4 Question Feature Extraction

The question features are the set of tokens generated by the BERT-Tokenizer. For tokenizing the question, a vocabulary is initially built with the text data available in the dataset. This vocabulary file is used with BERT to tokenize the question which is the required question feature. The Question Feature Extraction Process is explained in Algorithm 4.

Algorithm 4 Question Feature Encoding

Input : *Question* ▷ Input Question for Feature Extraction

Output : *Question_Encoding*

Tokenizer ← *BERT_Tokenizer*(*VocabFilePath*) ▷ Question Tokens

function GETTOKENIZEDQUESTION(Question)

Preprocessed_Question ← *textPreProcess*(*Question*)

Question_Encoding ← *Tokenizer*(*Preprocessed_Question*)

return *Question_Encoding*

end function

3.3 VQA MODEL BUILDING

The image and question features are extracted using VGGNet and BERT. These features are then fused by concatenating the feature vectors. VQA model building involves developing an answer generating model that takes the encoded image features and the tokenized question features as input and generates the corresponding answers. In this project, a BERT Model is used for generating answers by taking the fused features as input. A BERT model is very efficient for the tasks of Next Sentence Prediction (NSP) and Masked Language Modeling (MLM). MLM involves predicting the masked token in the given sentence or a paragraph. The idea of MLM is used in this project, to generate the answers for the given image and question using the fused feature vectors.

Training the BERT using MLM involves constructing the input for BERT with the image and text encoding along with the respective tokenized answers. The parts of the answers are masked and the model is trained to predict the masked words. The input for training is first constructed in the following format as explained by Algorithm 5:

[CLS] ENCODED-IMAGE-FEATURES [SEP] QUESTION-FEATURE [SEP] MASKED-ANSWER [SEP]

Algorithm 5 Constructing Input for BERT Training

Input: Image, Question, Answer

Output: Tokens ▷ A list of token vector

function CONSTRUCTINPUT(Image,Question,Answer)

$MaxLen \leftarrow 1000$

$ImageEncoding \leftarrow getImageEncoding(Image)$

$QuestionEncoding \leftarrow getTokenizedQuestion(Question)$

$AnswerEncoding \leftarrow getTokenizedQuestion(Answer)$

$Input \leftarrow [CLS] + ImageEncoding + [SEP] + QuestionEncoding + [SEP] + AnswerEncoding + [SEP]$

$Padding \leftarrow MaxLen - len(Input)$

$i \leftarrow 0$

$Tokens \leftarrow []$ ▷ Empty List

while $i < Padding$ **do**

$temp\_tokens \leftarrow Input$

$mask\_positions \leftarrow len(QuestionEncoding + ImageEncoding) + 3 + i$

$temp\_tokens[mask\_positions] \leftarrow MASK$ ▷ Masking

$Tokens.append(temp\_tokens)$

$i \leftarrow i + 1$

end while

return *Tokens*

end function

Figure 3.4 shows the working of the BERT model to predict the masked word.

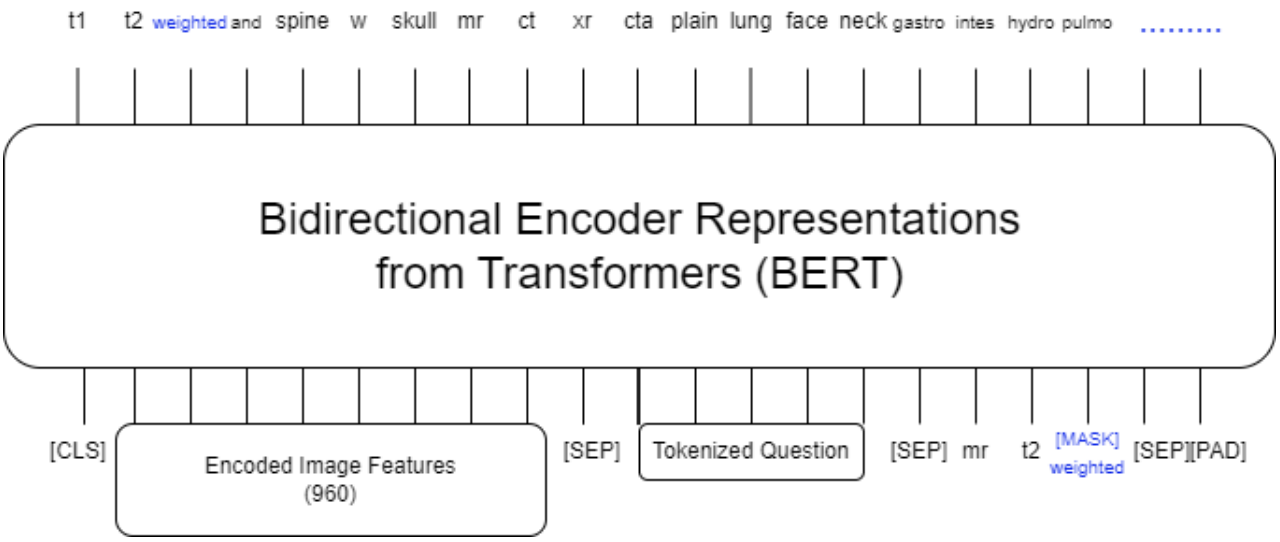


FIGURE 3.4: BERT model predicting the masked word

The trained model is then used for generating answers. For generating answers from the trained model, the input data is of the form:

[CLS] ENCODED-IMAGE-FEATURES [SEP] QUESTION-FEATURE [SEP] [MASK]

The model now attempts to predict the word at the masked position. When the model predicts the word, then the word is concatenated to the input data and the [MASK] is appended to the end of it. Now the model tries to predict the word at the current position of [MASK]. The above process repeats until a [SEP] token is predicted, marking the end of the answer. This is summarized in the Algorithm 6.

Algorithm 6 Answer Generation

Input: Image, Question

Output: Answer

function GETANSWER(Image,Question)

$MaxLen \leftarrow 1000$

$ImageEncoding \leftarrow getImageEncoding(Image)$

$QuestionEncoding \leftarrow getTokenizedQuestion(Question)$

$Input \leftarrow [CLS] + ImageEncoding + [SEP] + QuestionEncoding + [SEP] + [MASK]$

$Padding \leftarrow MaxLen - len(Input)$

$i \leftarrow 0$

$Vocab \leftarrow loadVocabulary(VocabPath)$

$Answer \leftarrow$ "" ▷ Empty String

$MaskPosition \leftarrow len(Input)$

while $i < Padding$ **do**

$Prediction \leftarrow VQAModel(Input)$

$GeneratedWord \leftarrow Vocab[Prediction]$

if $GeneratedWord =$ "$[SEP]$" **then**

break

end if

$Answer \leftarrow Answer + GeneratedWord$

$Input[MaskPosition] \leftarrow Prediction$

$MaskPosition \leftarrow MaskPosition + 1$

$Input[MaskPosition] \leftarrow [MASK]$

$i \leftarrow i + 1$

end while

return *Answer*

end function

For instance, to generate the answer **'bucket handle tear of meniscus'** (Image ID: synpic58267), the model generates answer as shown in the Figure 3.5

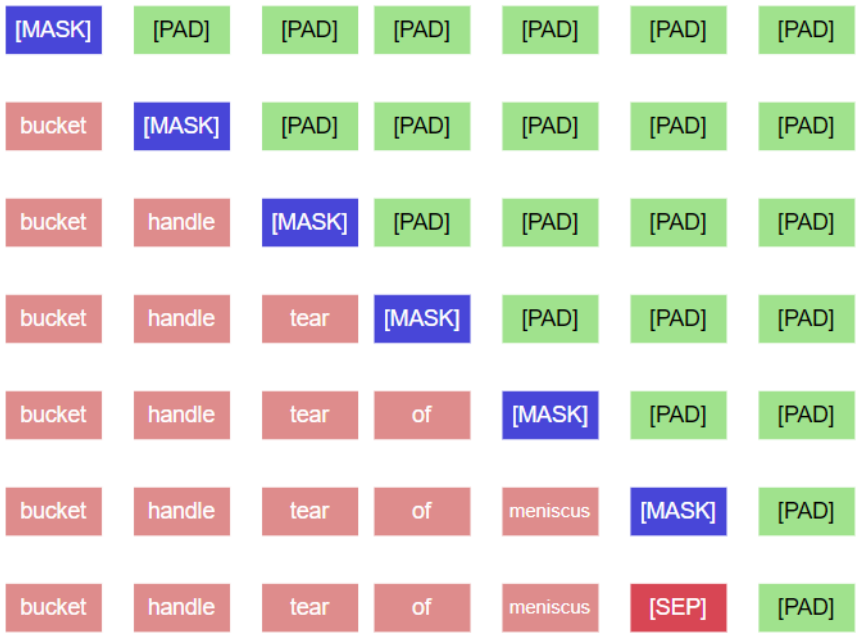


FIGURE 3.5: Answer Generation for a Sample with ID: synpic58267

3.4 EXPLAINABLE AI TECHNIQUES

Explainable AI (XAI) deals with explaining the predictions made by a ML/DL model. They try to analyze the output with respect to the input features that contributed to the arrival of that prediction. There are many XAI techniques like LIME, SHAP, Anchor, Contextual Importance and Utility (CIU), Gradient-weighted Class Activation Mapping (Grad-CAM) and etc., to interpret various ML/DL models and give explanations for the output. In this project LIME and SHAP are used for analyzing the outcome and provide explanations.

3.4.1 XAI - LIME

Local Interpretable Model-agnostic Explanations (LIME) [14] builds an interpretable local model that is trained on a perturbed dataset generated from the given input data samples. The local model is trained to approximate the predictions of the actual model to be interpreted.

The explanation is defined as a model $g \in G$, where G is class of interpretable models such as linear models or decision trees or falling rule lists (if-then rules). As not every $g \in G$ may be simple enough to be interpretable, the complexity $\Omega(g)$ is taken into account while finding the explanations. The complexity $\Omega(g)$ can be depth of the tree (in case of decision tree) or the number of non-zero weights (in case of linear models).

Let f denotes the model being explained and $f(x)$ is the probability that x belongs to a certain class. where x is the instance to be explained. Further $\pi_x(z)$ is the proximity measure between instances x and z. Finally, let $\mathcal{L}(f, g, \pi_x)$ be a measure of how unfaithful g is in approximating f in the locality defined by π_x.

The explanation ξ, for an instance x produced by LIME is obtained using the Eq. 3.1

$$\xi(x) = \underset{g \in G}{\arg\min}\ \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{3.1}$$

LIME generates explanations by training a local model on the perturbed dataset. Perturbed dataset is a dataset that is created by modifying the incoming input instance. In this project, LIME Image Explainer is used for generating explanations. LIME image explainer accepts a prediction function which accepts the input for the model, a instance to explain, inpaint color (shades of gray) ie.,

hide color and number of samples in the perturbed dataset. The perturbed dataset is generated by randomly inpainting segments of the input. A local model is trained on this perturbed dataset. The local model is an interpretable model. This local model is interpreted and analyzed for finding the segments of the image that affect the output towards predicting the same outcome as that of the actual model to be interpreted.

3.4.2 XAI - SHAP

SHapley Additive exPlanations (SHAP) [18] is an XAI framework that is derived from game theory. It works based on Game Theory proposed by mathematician John von Neumann and economist Oskar Morgenstern in the 1940s. Game Theory is the study of how the participation of players (in our case Features) influence the outcome. Later Lloyd Shapley introduced a measure to fairly distribute both gains and costs to several players (in our case features) working in coalition (working as a team). In honour of Lloyd Shapley, this measure is named as Shapley Values [27].

The aim of SHAP is to explain the prediction or the outcome of an instance x by computing the contribution of each feature of the instance x to the prediction [28]. SHAP calculates the Shapley Values from Game Theory. A feature value can act alone or as a group to arrive at an outcome. The Shapley values specifies how to fairly distribute the gain and the costs among the group of feature values. The Shapley value is represented as an additive feature attribution method as described by Eq. 3.2.

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j^{'} \tag{3.2}$$

where g is the explanation model, $z' \in \{0,1\}^M$ is the coalition vector, M is the maximum coalition size and $\phi_j \in \mathbb{R}$ is the feature attribution for a feature j, the Shapley values.

In this project, SHAP Partition Explainer is used for generating explanations. SHAP Partition Explainer recursively computes the Shapley values for hierarchical combinations of features. It captures the relationship between a combination of related features. In SHAP Partition Explainer, a masker is used that masks the regions of the image and finds the impact of masking a region. The impact is computed in terms of Shapley Values. The Shapley values of each features of input are calculated and are visualized using SHAP's Image Plot.

3.5 PERFORMANCE ANALYSIS USING QUANTITATIVE METRICS

The performance of Visual Question Answering Models can be analyzed using various metrics such as Accuracy, Bilingual Evaluation Understudy (BLEU) Score and Word-based Semantic Similarity (WBSS). Accuracy is a measure of how accurate the generated answer matches with the actual answer. The perfect match gives score 1, otherwise 0. Accuracy is calculated as shown in Eq. 3.3.

$$Accuracy = \frac{No.\ of\ correctly\ generated\ answers}{Total\ no.\ of\ samples} \tag{3.3}$$

The BLEU Score or the Bilingual Evaluation Understudy Score is a score for comparison of the generated answer and the actual answers. The comparison here does not check if both the answers exactly match. It calculates the score based on how much of the answer words match in each of the generated and actual answers. The perfect match gives a BLEU score of 1.0 while a perfect mismatch gives 0. The BLEU score value can be between 0 and 1, depending on the percentage of match between the two answers. The steps involved in calculating BLEU Score are as follows.

The first step is to compute Precision scores for 1-grams. The Precision scores for 1-grams is calculated as given by Eq. 3.4

$$PrecisionOneGram = \frac{No.\ of\ CorrectlyPredictedOneGrams}{No.\ of\ TotalPredictedOneGrams} \tag{3.4}$$

The next step is to calculate the Brevity Penalty using the values of c, which is the number of words in the generated sentence and r, which is the number of words in the target sentence. The Brevity Penalty is calculated as shown in Eq. 3.5.

$$BrevityPenalty = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c <= r \end{cases} \tag{3.5}$$

Finally to calculate BLEU Score, the Brevity Penalty is multiplied with Precision 1-gram value as given by Eq. 3.6

$$BLEU = BrevityPenalty\ .\ PrecisionOneGram \tag{3.6}$$

Word-based Semantic Similarity (WBSS) is used to compare the Wu-Palmer similarity (WUPS) between the words in each of the actual and generated answers. For a generated answer A and the actual answer or the ground truth T, the WUPS[15] is calculated as depicted in Eq. 3.7.

$$WUPS(A,T) = \frac{1}{N} \times \sum_{i=1}^{N} \times min\left\{ \prod_{a \in A^i} max_{t \in T^i} WUP(a,t), \prod_{t \in T^i} max_{a \in A^i} WUP(a,t) \right\} \times 100 \tag{3.7}$$

3.6 IMPLEMENTATION FILES

1. **File Name:** *VQA_Model_Training.ipynb* (IPython Notebook)
   Input: Train and Validation Datasets
   Output: Trained VQA Model
   Description: This IPython Notebook has functions for Training the VQA Model using VGGNet for Image Feature Extraction, BERT Tokenizer for tokenizing the Question. The features are concatenated and a BERT Model is trained for Answer Generation using the concatenated features.

2. **File Name:** *Testing_and_Evalutaion.ipynb* (IPython Notebook)
   Input: Trained VQA Model and Test Dataset
   Output: Performance Metrics
   Description: This notebook tests the trained VQA Model with the Test Dataset and evaluates it based on various performance metrics such as Accuracy, BLEU Score and WBSS.

3. **File Name:** *VQA_and_XAI.ipynb* (IPython Notebook)

   Input: Trained VQA Model and set of images & quesions

   Output: Explanations

   Description: This notebook uses the Explainable AI technique - SHAP to provide explanations to the output of the Trained VQA Model.

CHAPTER 4

RESULTS AND PERFORMANCE ANALYSIS

In this chapter, the results of every stage of the project from Dataset Collection & Analysis, Feature Extraction, Visual Question Answering (VQA) Model Building, Model Interpretation using Explainable AI (XAI) and Performance Analysis using Quantitative Metrics are explained with required snapshots and analysis.

4.1 RESULT OF DATASET ANALYSIS

Many Datasets are available for the desired tasks and they contain many Medical Images along with many relevant questions for each image. A ImageCLEF task has been posted for VQA for Medical Images and the dataset (VQA-Med 2019) is under AI Crowd. The dataset contains different medical images and their corresponding Question-Answer pairs. There are 3200 training medical images. For each image, there are four questions. The questions are categorized into four major types - Modality, Plane, Organ System and Abnormality. There are totally 12,792 Question-Answer pairs.

A text file for each category of questions is given in the dataset which includes the Question-Answer pair along with the image ID which corresponds to the name of the image file.

The validation set contains 500 images and 2000 Question-Answer pairs. The test set consists of 500 images and 500 Question-Answer pairs. The result of the analysis is given in the Table 4.1.

Dataset	Images	Question Category	No. of Questions	No. of Classes
Training	3200	Modality	3200	44
		Plane	3200	15
		Organ System	3200	10
		Abnormality	3192	1484
		Total	12792	-
Validation	500	Modality	500	-
		Plane	500	-
		Organ System	500	-
		Abnormality	500	-
		Total	2000	-
Testing	500	All	500	-

TABLE 4.1: Dataset Analysis

Figure 4.1 shows a sample of a single image with a question and the corresponding answer.

FIGURE 4.1: A Sample Image and a QA pair from Dataset (Image ID: synpic52980)

Figure 4.2 shows a sample of a single image with 4 questions and the corresponding answers.

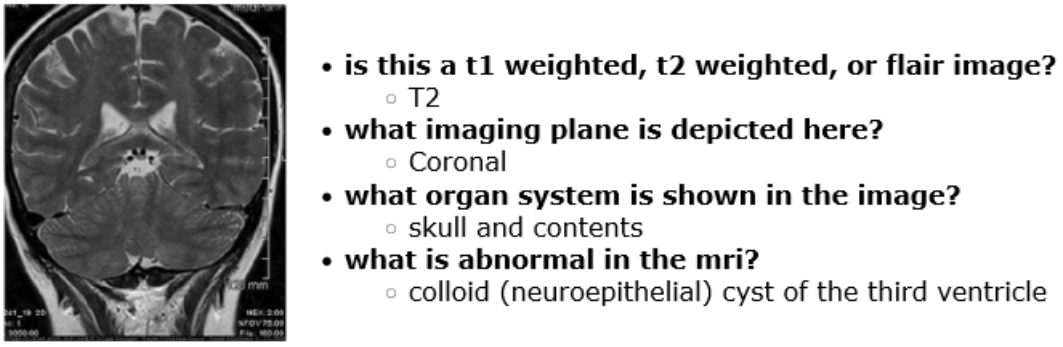


FIGURE 4.2: A Sample Image and 4 QA pairs from Dataset (Image ID: synpic16994)

Figure 4.3 shows a sample of 4 images with a common question and the corresponding answers.

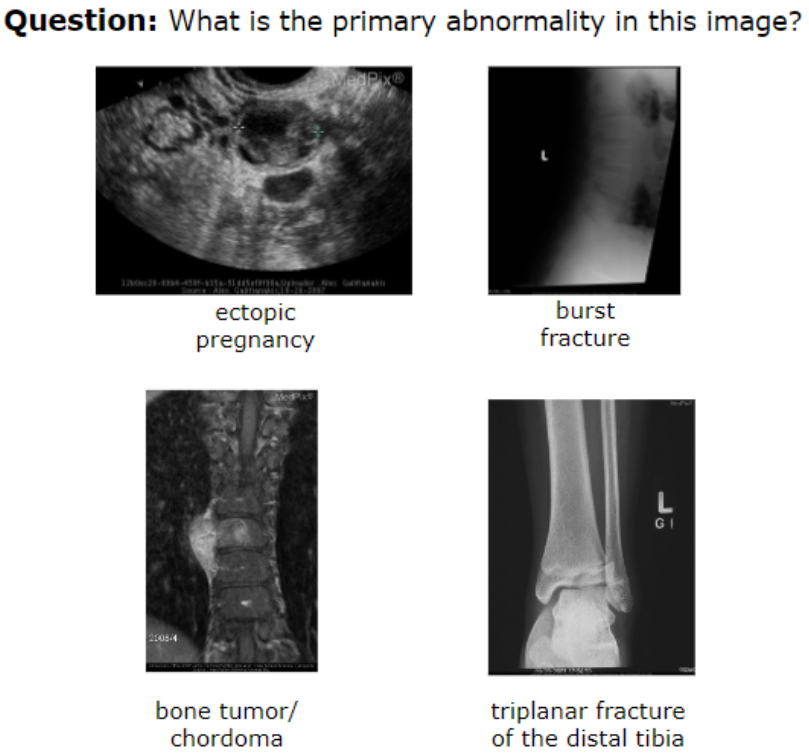


FIGURE 4.3: Sample Images with common questions (Image IDs: synpic38930, synpic52143, synpic20934, synpic19141)

4.2 RESULT OF FEATURE EXTRACTION

The images and questions are pre-processed and the features are extracted using VGGNet and Bilingual Evaluation Understudy (BERT) respectively.

4.2.1 Result of Image Feature Extraction

The image features are extracted and are encoded as explained in Algorithm 2 from the VGGNet model. The efficiency of various Convolutional Neural Network (CNN) models like a custom CNN, the pre-trained VGGNet and the VGGNet architecture trained using the organ dataset in extracting features from the images is analyzed using a heatmap which depicts the activations of the last Convolution Layer.

Figure 4.4 shows the generated heatmap of the activations in the last convolution layer of the custom CNN, superimposed onto the original image.

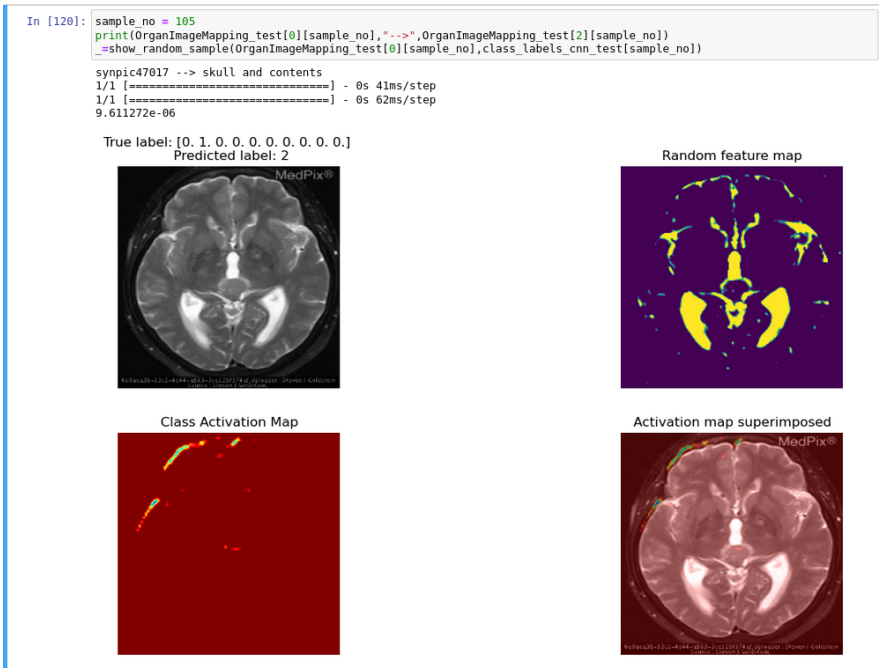


FIGURE 4.4: Activations of Custom CNN

Figure 4.5 shows the generated heatmap of the activations in the last convolution layer of the pre-trained VGGNet, superimposed onto the original image.

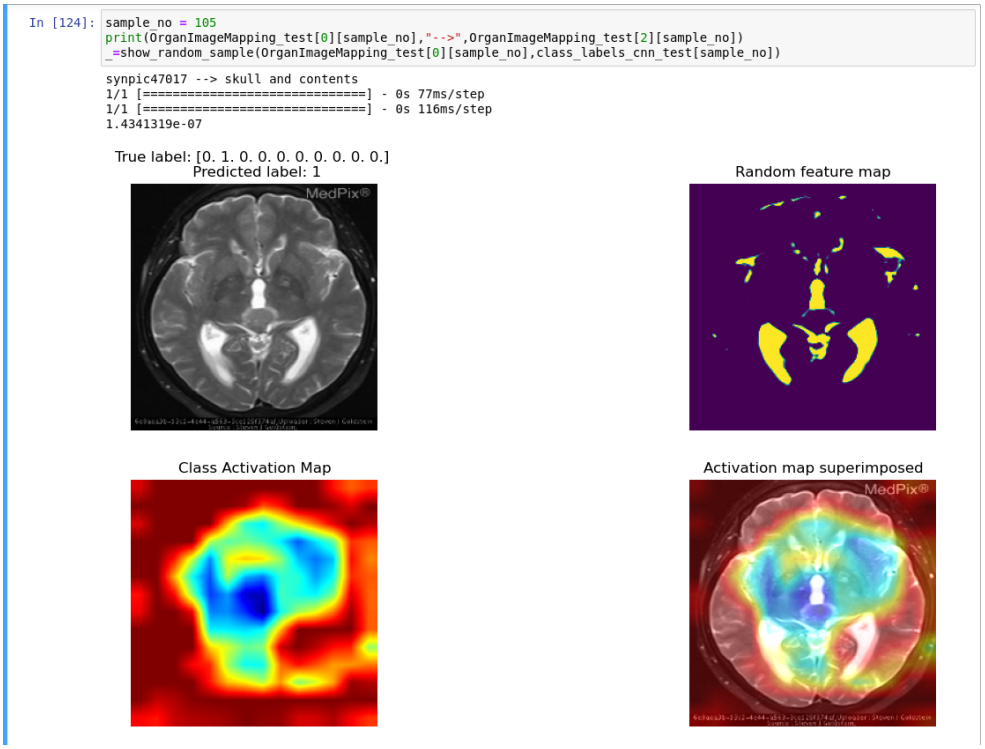


FIGURE 4.5: Activations of Pre-trained VGGNet

Figure 4.6 shows the generated heatmap of the activations in the last convolution layer of the VGGNet trained on Organ dataset, superimposed onto the original image.

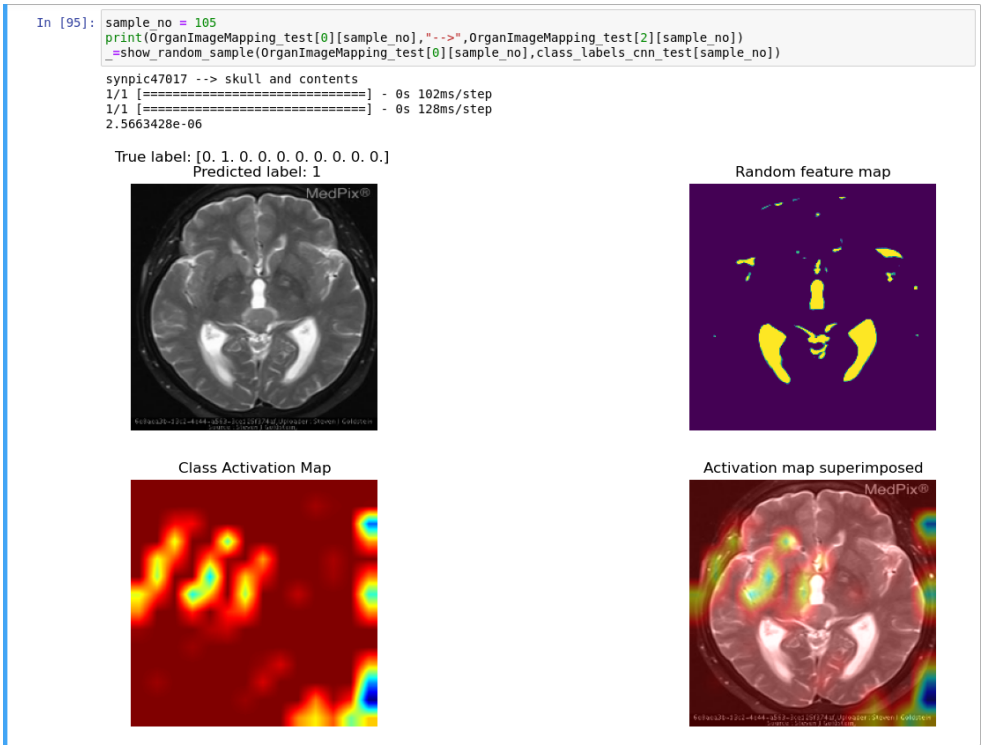


FIGURE 4.6: Activations of VGGNet trained on the organ dataset

From the above Activation heatmaps, it is evident that the pre-trained VGGNet can efficiently extract features from the images compared to the other two CNN architectures.

4.2.2 Result of Question Feature Extraction

The questions are tokenized using BERT-Tokenizer. To tokenize the question, a vocabulary is built using the text data available in the dataset. This vocabulary is

used to tokenize the question using BERT. The tokenized questions is the required question feature.

The vocabulary built using the text from the dataset has 4914 words including the special tokens for BERT such as [PAD], [UNK], [CLS], [SEP] and [MASK].

A sample set of tokens and their corresponding token ID is shown in the Table 4.2.

Token ID	**Token**
0	[PAD]
1	[UNK]
2	[CLS]
3	[SEP]
4	[MASK]
92	th
94	##at
97	what

TABLE 4.2: A sample set of tokens and their IDs from the vocabulary

4.3 RESULT OF VQA MODEL BUILDING

The feature from images and questions are extracted and are fused. For answer generation, a BERT model is built and trained for generating answer tokens. The model is trained for 20 epochs with batch size of 40. Figure 4.7 shows the training and validation of the model.

```
Epoch: 1

Iter (loss=0.065): : 1042it [21:09,  1.22s/it]
Iter (loss=0.802): : 176it [02:29,  1.18it/s]

Epoch: 2

Iter (loss=0.023): : 1042it [21:18,  1.23s/it]
Iter (loss=0.902): : 176it [02:28,  1.19it/s]

Epoch: 3

Iter (loss=0.014): : 1042it [21:20,  1.23s/it]
Iter (loss=0.752): : 176it [02:27,  1.19it/s]

Epoch: 4

Iter (loss=0.006): : 1042it [21:09,  1.22s/it]
Iter (loss=0.870): : 176it [02:28,  1.19it/s]

Epoch: 5

Iter (loss=0.007): : 1042it [21:21,  1.23s/it]
Iter (loss=0.631): : 176it [02:27,  1.19it/s]

Epoch: 6

Iter (loss=0.004): : 1042it [21:07,  1.22s/it]
Iter (loss=0.455): : 176it [02:28,  1.18it/s]

Epoch: 7

Iter (loss=0.002): : 1042it [21:10,  1.22s/it]
Iter (loss=2.232): : 176it [02:27,  1.20it/s]
```

FIGURE 4.7: Training and Validation of Model

The trained model is now used for generating answers for the given question along with the corresponding input image. Figures 4.8 - 4.11 shows samples of answers generated by the model.

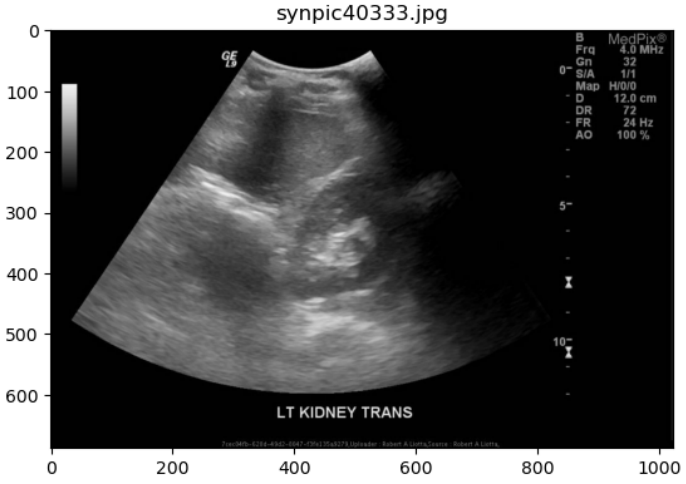
Figure 4.8 shows a sample image queried about the Modality.

Figure 4.9 shows a sample image queried about the Organ.

Figure 4.10 shows a sample image queried about the Plane.

Figure 4.11 shows a sample image queried about the Abnormality.

```
generateAnswer('synpic40333','what imaging modality was used to take this image?')
```

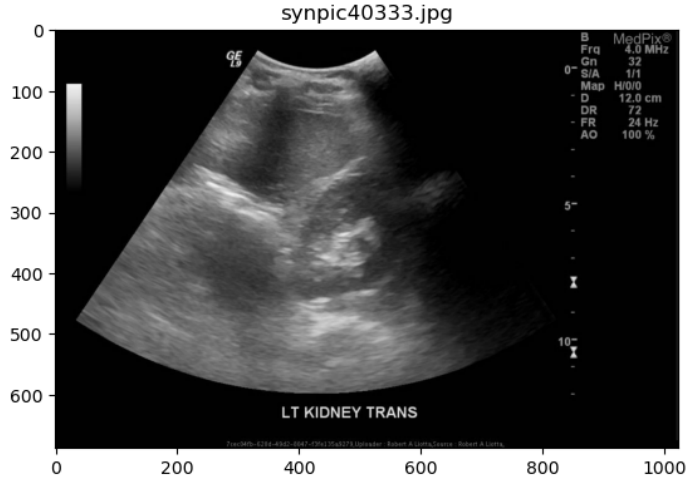



```
Generating Answers...
us
ultrasound
[SEP]

'us ultrasound '
```

FIGURE 4.8: Sample Answer Generation for the image with ID: synpic40333 and queried about the Modality (Correct Answer)

```
generateAnswer('synpic40333','what organ system is imaged?')
```

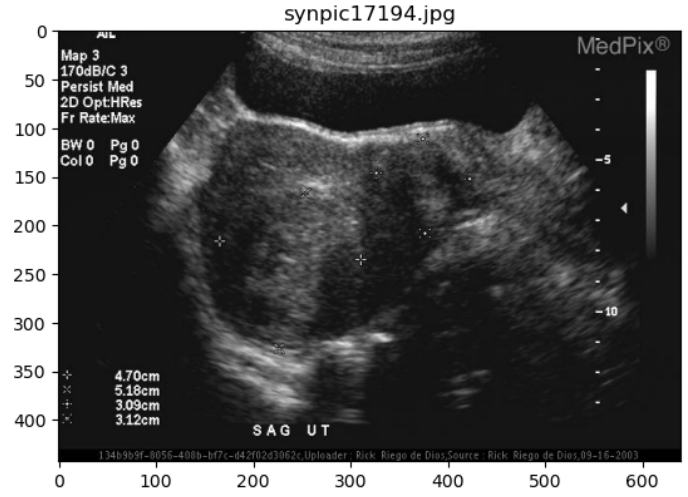



```
Generating Answers...
gastrointestinal
[SEP]

'gastrointestinal '
```

FIGURE 4.9: Sample Answer Generation for the image with ID: synpic40333 and queried about the Organ captured (Correct Answer)

```
generateAnswer('synpic17194','what is the plane shown in this image?')
```

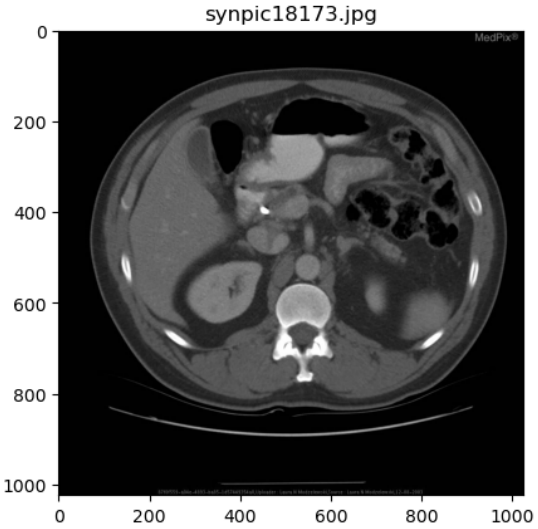



```
Generating Answers...
sagittal
[SEP]

'sagittal '
```

FIGURE 4.10: Sample Answer Generation for the image with ID: synpic17194 and queried about the Plane (Correct Answer)

```
generateAnswer('synpic18173','what is the primary abnormality in this image')
```




```
Generating Answers...
pancreatic
adenocarcinoma
[SEP]

'pancreatic adenocarcinoma '
```

FIGURE 4.11: Sample Answer Generation for the image with ID: synpic18173 and queried about the Abnormality (Wrong answer - Actual answer: pancreatic duct adenocarcinoma)

4.4 RESULT OF EXPLAINABLE AI

The outcome of the VQA model is analyzed using Explainable AI (XAI) techniques like Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). The result of the analysis gives the explanations for the outcome.

4.4.1 XAI - LIME

LIME works by building a local model and training it on a perturbed dataset created using the input instance. For example, for the sample image with ID: synpic56918 (Figure 4.12), the perturbed dataset generated of 5 samples (1 actual input image + 4 generated iages) is shown in Figure 4.13

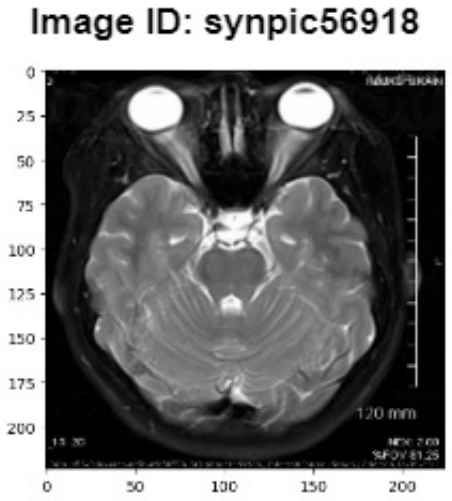


FIGURE 4.12: Sample input for LIME (Image ID: synpic56918)

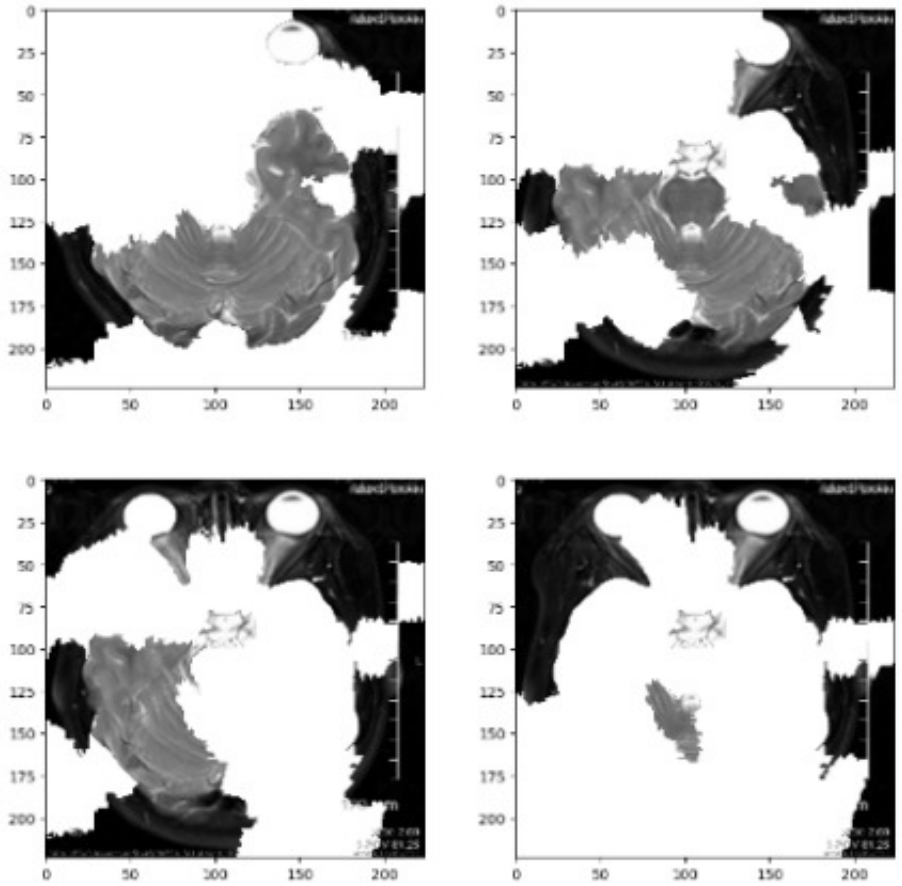


FIGURE 4.13: Perturbed Dataset for the sample input with ID: synpic56918

Figure 4.14 shows the LIME explanation for the sample image (Image ID: synpic56918) queried about the organ.

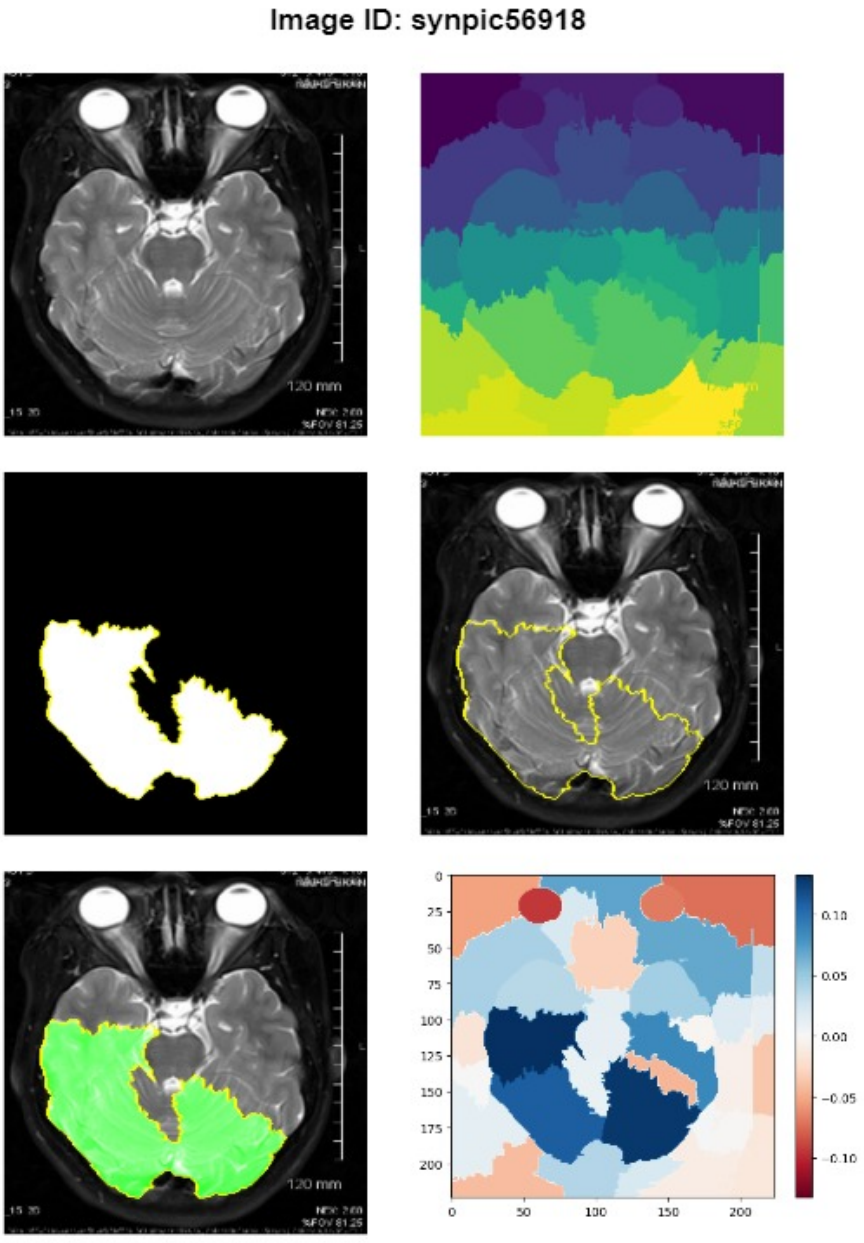


FIGURE 4.14: LIME explanations for a sample image with ID: synpic56918 queried about the organ

The green inpaint on the image shows the contributions of segments of the image which are positive to that of the outcome.

The following Figure 4.15 shows LIME explanations for 6 different input images with a common question about the organ captured by the image.

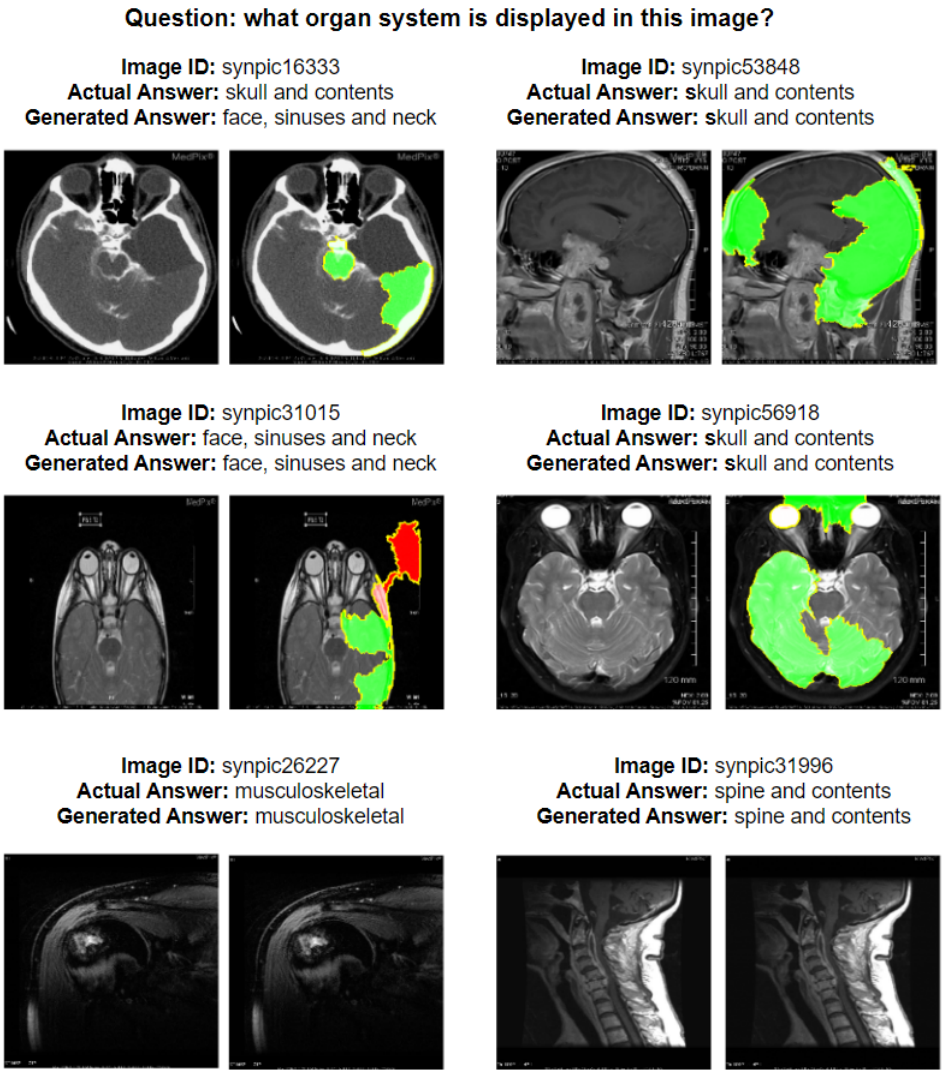


FIGURE 4.15: LIME explanations for a set of sample images queried about the organ

In the Figure 4.15, for the image with ID: synpic31015, there is red inpaint on the output image. The red inpaint indicates negative contribution which deviates the outcome from the actual model's outcome. It is also evident that for each input image the segments responsible for the prediction is different.

Figure 4.16 shows the LIME explanations for the sample image (Image ID: synpic56918) for 4 questions of different categories.

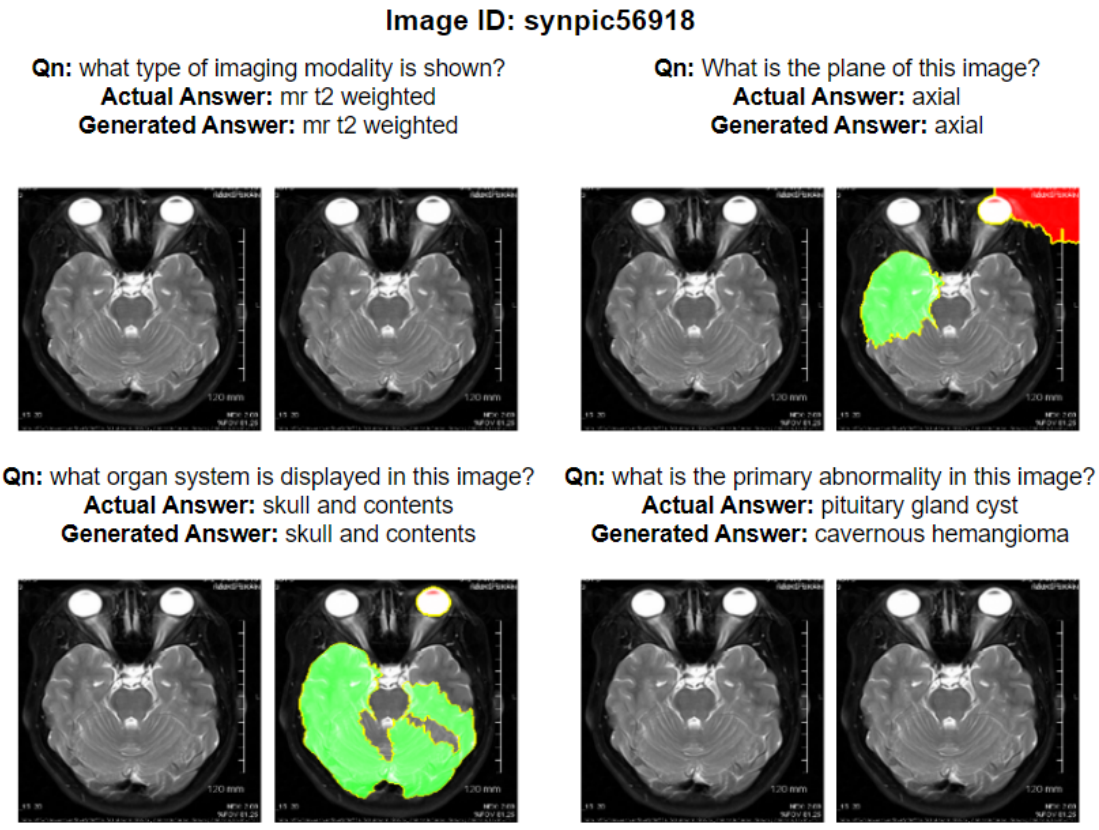


FIGURE 4.16: LIME explanations for a sample image with ID: synpic56918 for 4 different questions of different categories

From Figure 4.16, it is inferred that the parts of the image that contribute to the outcome is different for different questions which is indicated by the inpaint on the image. For some output images, there is no inpainting which indicates that LIME could not find necessary information for interpreting the outcome of the VQA model.

Figure 4.17 also shows the explanation generated for the second time for the same sample image (Image ID: synpic56918) queried about the organ.

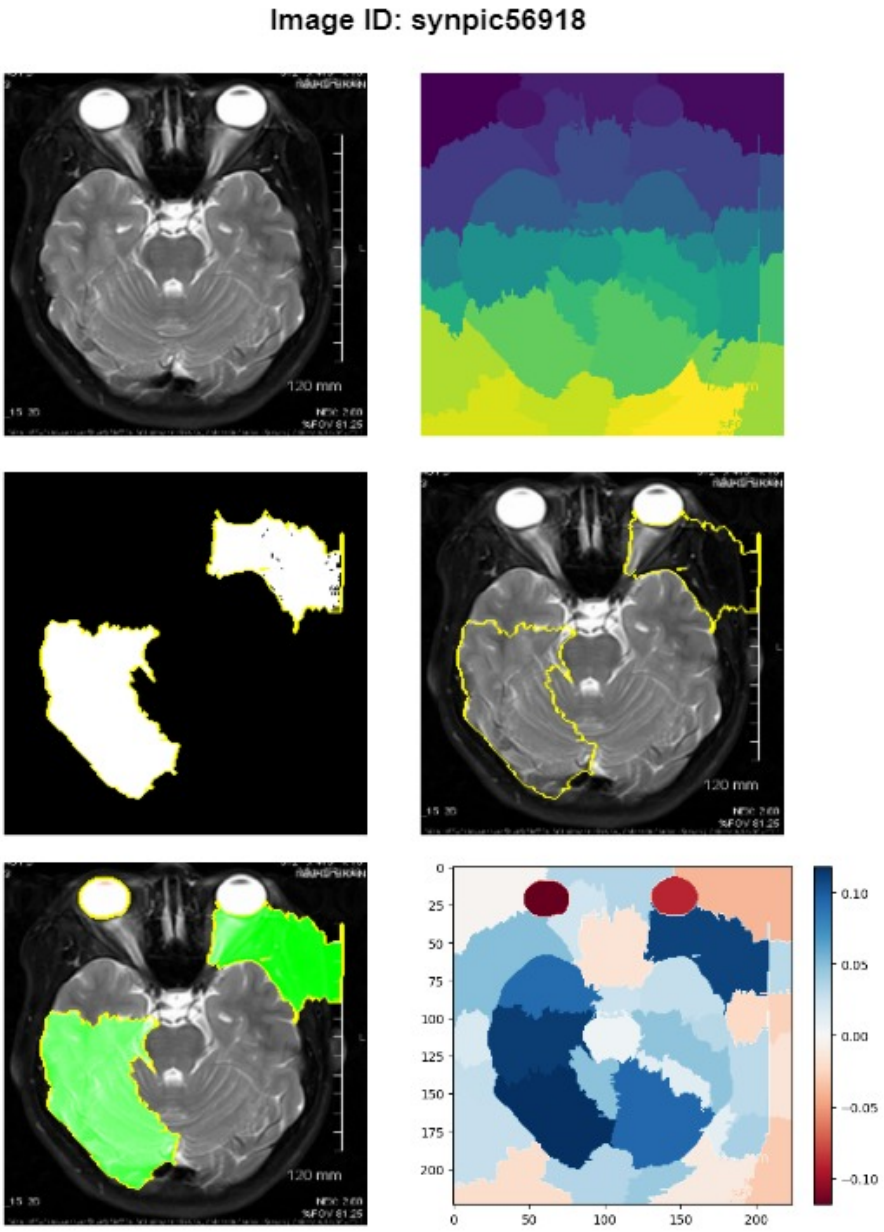


FIGURE 4.17: LIME explanations for a sample image with ID: synpic56918 queried about the organ on the second run

From the Figures 4.14 and 4.17, it is inferred that the explanations differ in each run. This is because of the perturbed dataset generated by LIME is different for every run. The explanations are not consistent and this has been overcome using SHAP which is based on Game Theory and Shapely values.

4.4.2 XAI - SHAP

SHAP calculates Shapely values for each features in the input instance. The Shapely values measures the contribution of the feature to the outcome. SHAP takes a single input instance or a set of input instances as input to generate explanations. The SHAP Partition Explainer is used to generate explanations in this project. SHAP Partition Explainer masks parts of the image and calculates the Shapely values. A SHAP Partition Explainer is initialized with the model to be interpreted, a masker and the class labels. Maskers are used to mask out the regions of the image by inpainting or blurring. SHAP uses these maskers and computes the Shapely values for the parts of the image. The computed Shapely values are then visualized using SHAP's Image Plot method.

Figure 4.18 shows shows the output of SHAP for the image with ID: synpic56918 queried about the organ system.

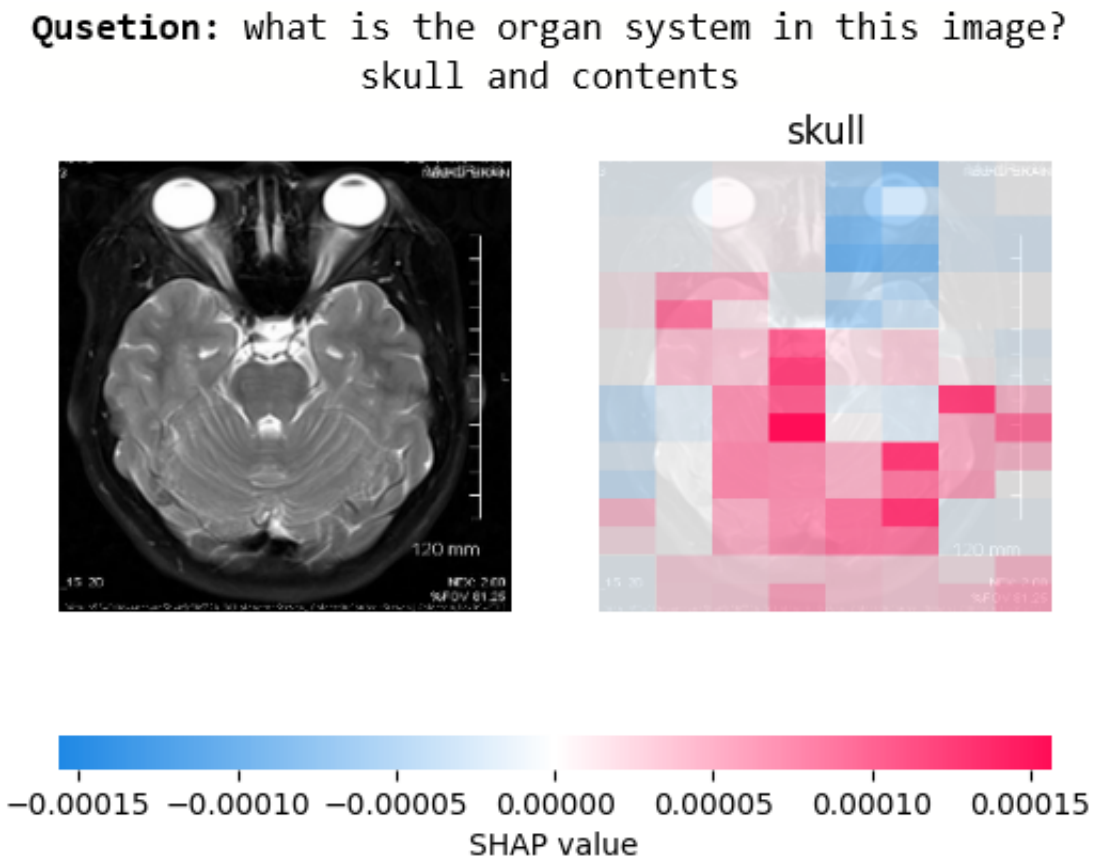


FIGURE 4.18: SHAP Explanations for a sample image with ID: synpic56918 queried about organ

The regions highlighted with red color represents the positive contribution to the output and regions with blue color represents negative contribution to the output. The white regions does not affect the output in any way.

From the Figure 4.18, it is inferred that the regions of the skull are colored mostly with various intensities of red. These regions contributed highly for the prediction of the organ "skull".

Figure 4.19 shows the SHAP explanations for the prediction of the organ system by the VQA model for different images with IDs: synpic16333, synpic53848, synpic31015, synpic56918, synpic26227 and synpic31996.

FIGURE 4.19: SHAP explanations for a set of sample images queried about the organ

To infer the contributions of regions of image for different types of questions i.e.,

to infer the model's working for different types of questions, SHAP is applied to the same image for different questions. Figure 4.20 shows the SHAP explanations for the image with ID: synpic56918 for different questions.

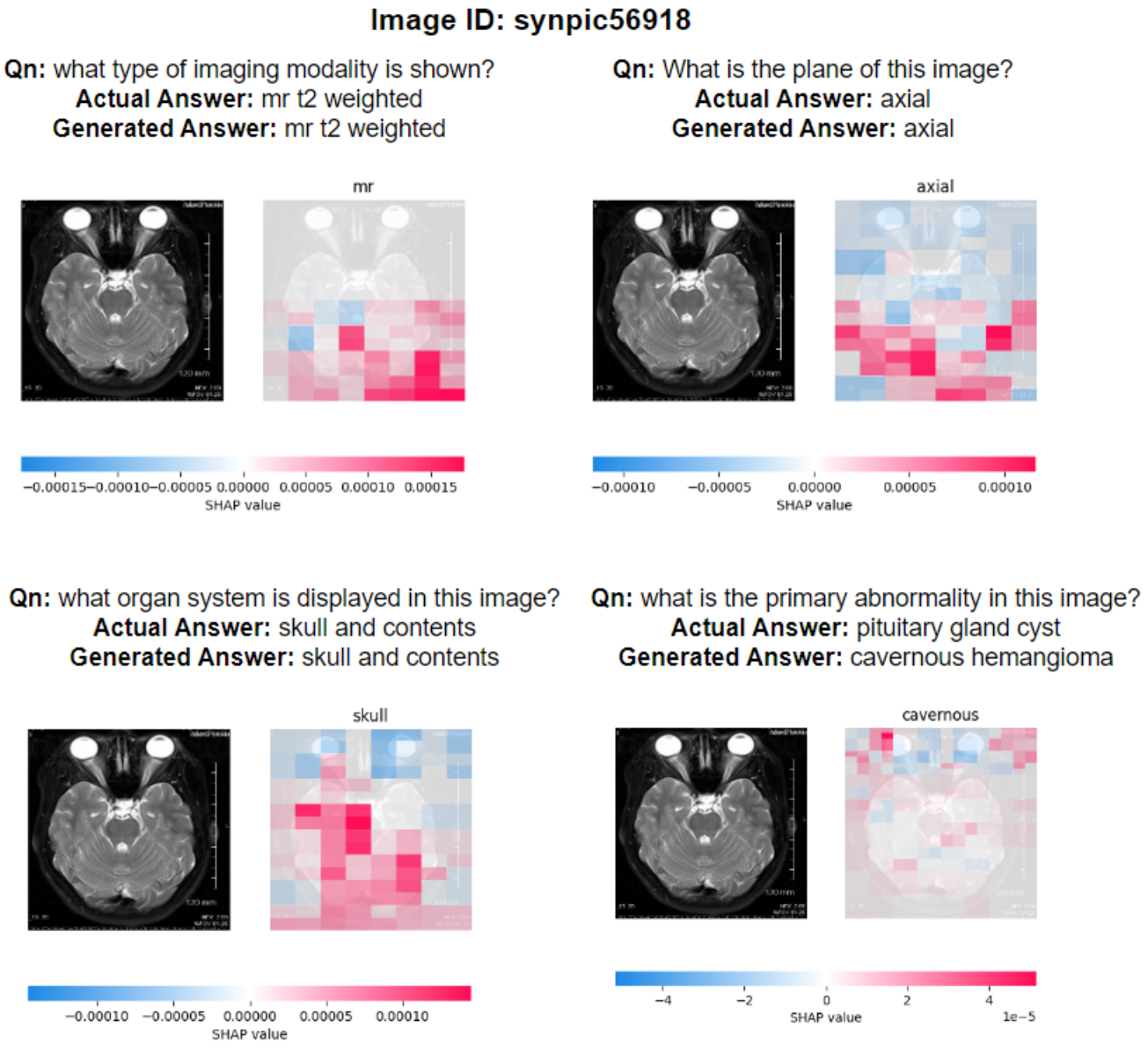


FIGURE 4.20: SHAP explanations for the model's working based on the question

4.5 RESULT OF PERFORMANCE ANALYSIS USING QUANTITATIVE METRICS

The performance of the VQA Model is analyzed using metrics such as accuracy, Bilingual Evaluation Understudy (BLEU) Score and Word-based Semantic Similarity (WBSS). Table 4.3 shows the performance of the VQA model for each categories and overall test data.

Category	**No. of Samples**	**Accuracy**	**BLEU Score**	**WBSS**
Modality	125	65.6	68.79	71.66
Plane	125	64.8	64.8	65.35
Organ	125	50.4	53.19	54.82
Abnormality	125	6.4	7.65	12.03
Overall	**500**	**46.8**	**48.61**	**50.97**

TABLE 4.3: Performance analysis using Accuracy, BLEU Score and WBSS

Table 4.4 show the comparison of the performance of the proposed model to the task winner [23]. The WBSS score was not calculated during the task and the WBSS score for the winner was not computed.

Model	**Accuracy**	**BLEU Score**	**WBSS**
Task Winner	62.4	64.4	–
Proposed Model	46.8	48.61	50.97

TABLE 4.4: Performance Comparison

The performance comparison summarized in Table 4.4 is visualized as a graph in Figure 4.21.

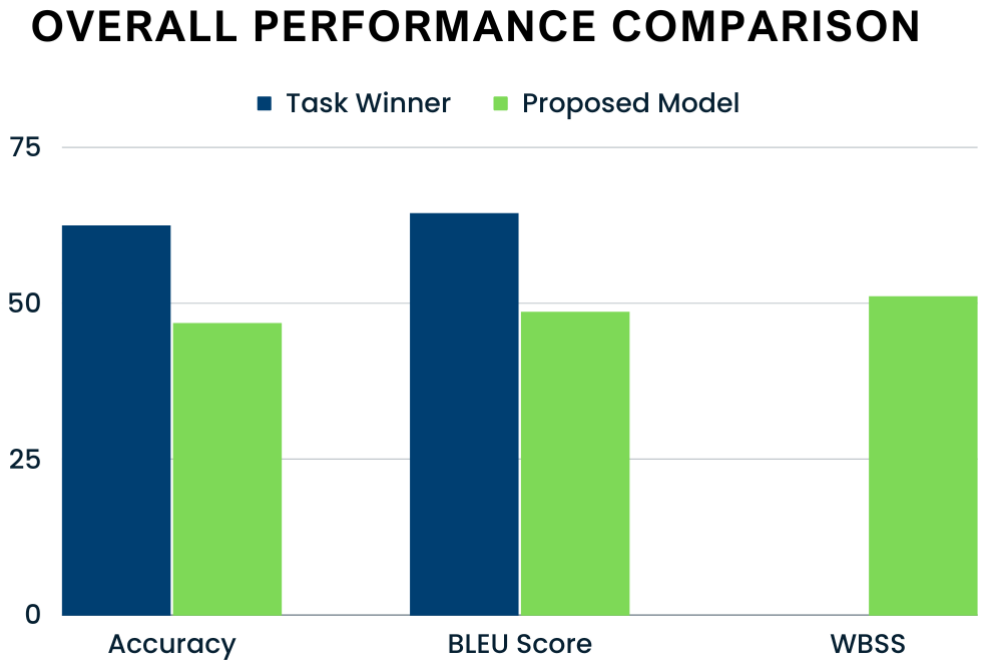


FIGURE 4.21: Overall Performance Comparison of Proposed Model with the Task Winner

Discussion

The performance metrics such as Accuracy, BLUE Score and WBSS of Modality based questions are high compared to other categories. In case of abnormality, the accuracy is very low due to unavailability of enough data for 1484 classes of abnormality in the training set (Refer Table 4.1 for analysis of the dataset). For other categories of questions, the classes are few and there is enough data for training.

CHAPTER 5

CONCLUSION AND FUTURE WORK

The dataset for the task of Visual Question Answering (VQA) is collected and it is analyzed in terms of the categories of the question. The image features are extracted using VGG-16 from the newly added dense layer of 960 units and are encoded. The question is tokenized using Bidirectional Encoder Representations from Transformers (BERT) with the help of the vocabulary built from the text data in the dataset. A BERT model is trained for predicting a masked token in the input consisting of concatenated image and question features. The trained model is then used to predict answers for the test data by recursively feeding the model with the encoded features and the answer from the last iteration. The model gives a 46.8% Accuracy, 48.61 BLEU Score and 50.97 WBSS for test dataset. The accuracy of abnormality based questions is too low due to data scarcity and it reduces the overall accuracy. Explainable AI (XAI) technique SHAP is used generate explanations for the predicted outcome and justify the outcome.

In future, an efficient Visual Question Answering model can be built for the ImageCLEF 2019 VQA-Med dataset. Also other XAI techniques could be explored by integrating with the VQA model for explanations.

REFERENCES

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

2. Abhishek Thanki and Krishnamoorthi Makkithaya, (2019) *MIT manipal at ImageCLef 2019 visual question answering in medical domain*. CEUR-WS.org - CLEF 2019 Working Notes, Vol. 2380.

3. Aisha Al-Sadi, Mahmoud Al-Ayyoub, Yaser Jararweh and Fumie Costen, (2021) *Visual question answering in the medical domain based on deep learning approaches: A comprehensive study*. Pattern Recognition Letters, Vol. 150 , pp. 57-75, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2021.07.002.

4. Aisha Al-Sadi, Talafha Bashar, Mahmoud Al-Ayyoub, Yaser Jararweh and Fumie Costen, (2019) *JUST at ImageCLEF 2019 Visual Question Answering in the Medical Domain*. CEUR-WS.org - CLEF 2019 Working Notes, Vol. 2380.

5. Avleen Malhi, Timotheus Kampik, Husanbir Pannu, Manik Madhikermi and Kary Främling, (2019) *Explaining Machine Learning-Based Classifications of In-Vivo Gastral Images*, 2019 Digital Image Computing: Techniques and Applications (DICTA), Perth, WA, Australia, pp. 1-7, https://doi.org/10.1109/DICTA47822.2019.8945986

6. Ben Abacha Asma, Hasan Sadid, Datla Vivek, Liu Joey , Demner-Fushman Dina and Müller Henning, (2019) *VQA-Med: Overview of the Medical Visual*

Question Answering Task at ImageCLEF 2019, CLEF 2019 Working Notes, CEUR Workshop Proceedings 2019.

7. Bounaama Rabia and Mohammed El Amine Abderrahim, (2019) *Tlemcen University at ImageCLEF 2019 Visual Question Answering Task*, CEUR-WS.org - CLEF 2019 Working Notes, Vol. 2380.

8. Cameron Severn, Krithika Suresh, Carsten Görg, Yoon Seong Choi, Rajan Jain and Debashis Ghosh, (2022) *A Pipeline for the Implementation and Visualization of Explainable Machine Learning for Medical Imaging Using Radiomics Features*. Sensors 22, Vol. 14, p. 5205, https://doi.org/doi:10.3390/s22145205

9. Dhruv Sharma, Snajay Purushotham and Chandan K Reddy, (2021) *MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain* Scientific Reports 11, Article No.: 19826. https://doi.org/10.1038/s41598-021-98390-1

10. Fuji Ren and Yangyang Zhou, (2020) *CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering*, in IEEE Access, Vol. 8, pp. 50626-50636, https://doi.org/10.1109/ACCESS.2020.2980024

11. Imane Allaouzi, Mohamed Ben Ahmed, Badr Benamrou, (2019) *An Encoder-Decoder Model for Visual Question Answering in the Medical Domain*, CEUR-WS.org - CLEF 2019 Working Notes, Vol. 2380.

12. Jannis Born, Nina Wiedemann, Manuel Cossio, Charlotte Buhre, Gabriel Brändle, Konstantin Leidermann, Julie Goulet, Avinash Aujayeb, Michael Moor, Bastian Rieck and Karsten Borgwardt, (2021). *Accelerating Detection of Lung Pathologies with Explainable Ultrasound Image Analysis*, Applied Sciences 11, Vol. 2, p. 672. https://doi.org/10.3390/app11020672

13. Knapič Samanta, Avleen Malhi, Rohit Saluja and Kary Främling, (2021) *Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain*. Machine Learning and Knowledge Extraction, Vol. 3, pp. 740-770. https://doi.org/10.3390/make3030037.

14. Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin, (2016) *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 97–101, San Diego, California. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-3020

15. Mateusz Malinowski and Mario Fritz, (2014) *A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input*. Adv Neural Inf Process. https://doi.org/10.48550/arXiv.1410.0210

16. Minh H. Vu, Raphael Sznitman, Tufve Nyholm and Tommy Löfstedt , (2019) *Ensemble of Streamlined Bilinear Visual Question Answering Models for the ImageCLEF 2019 Challenge in the Medical Domain*. CEUR-WS.org - CLEF 2019 Working Notes, Vol. 2380.

17. Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh and Dhruv Batra, (2020) *Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization*. International Journal of Computer Vision Vol. 128, pp. 336–359. https://doi.org/10.1007/s11263-019-01228-7

18. Scott M. Lundberg and Su-In Lee, (2017) *A Unified Approach to Interpreting Model Predictions*, Advances in Neural Information Processing Systems 30, pp. 4765–4774.

19. Shengyan Liu, Xiaozhi Ou, Jiao Che, Xiaobing Zhou and Haiyan Ding, (2019) *An Xception-GRU Model for Visual Question Answering in the Medical Domain*. CEUR-WS.org - CLEF 2019 Working Notes, Vol. 2380.

20. Shi Lei, Feifan Liu, and Max P. Rosen, (2019) *Deep Multimodal Learning for Medical Visual Question Answering*. CEUR-WS.org - CLEF 2019 Working Notes, Vol. 2380.

21. Sudil Hasitha Piyath Abeyagunasekera, Yuvin Perera, Kenneth Chamara, Udari Kaushalya, Prasanna Sumathipala and Oshada Senaweera, (2022) *LISA : Enhance the explainability of medical images unifying current XAI techniques*, IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, pp. 1-9, https://doi.org/10.1109/I2CT54291.2022.9824840

22. Urja Pawar, Donna O'Shea, Susan Rea and Ruairi O'Reilly, (2020) *Explainable AI in Healthcare*, International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), Dublin, Ireland, pp. 1-2, https://doi.org/10.1109/CyberSA49311.2020.9139655.

23. Xin Yan, Lin Li, Chulin Xie, Jun Xiao and Lin Gu, (2019) *Zhejiang University at ImageCLEF 2019 Visual Question Answering in the Medical Domain*, CEUR-WS.org - CLEF 2019 Working Notes, Vol. 2380.

24. Yangyang Zhou, Xin Kang and Fuji Ren, (2019) *TUA1 at ImageCLEF 2019 VQA-Med: a Classification and Generation Model based on Transfer Learning*. CEUR-WS.org - CLEF 2019 Working Notes, Vol. 2380.

25. Yu-Huan Wu, Shang-Hua Gao, Jie Mie, Jun Xu, Deng-Ping Fan, Rong-Guo Zhang and Ming-Ming Cheng, (2021) *JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation*, in IEEE Transactions on Image Processing, Vol. 30, pp. 3113-3126, 2021, https://doi.org/10.1109/TIP.2021.3058783

26. Zhihong Lin, Donghao Zhang, Qingyi Tac, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He and Zongyuan Ge, (2021) *Medical visual question answering: A survey*, arXiv preprint arXiv:2111.10056, https://doi.org/10.48550/arXiv.2111.10056.

27. Shapley Value : https://www.investopedia.com/terms/s/shapley-value.asp, April 2023

28. SHAP : https://christophm.github.io/interpretable-ml-book/shap.html, April 2023.