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	**DR. S. KAVITHA**
HEAD OF THE DEPARTMENT	**SUPERVISOR**
Professor,	Associate Professor,
Department of CSE,	Department of CSE,
SSN College of Engineering,	SSN College of Engineering,
Kalavakkam - 603 110	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. This project is proposed to develop a VQA model for ImageClef 2019 VQA-Med Dataset and the predicted outcome is analyzed using Explainable AI (XAI) techniques. The images of the chosen dataset are complex to analyze and are of low resolution with multiple questions for each image. 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 a test dataset and 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 XAI techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). The result of this analysis gives the explanations on the outcome. The explanations reveal how and why a prediction has been made and justify the predicted outcome.

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
GAP	Global Average Pooling

CHAPTER 1

INTRODUCTION

Artificial Intelligence (AI) has grown exponentially over the past 10-15 years. The intelligent models or agents developed using AI have solved many real-world problems. 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 Natural Language Processing and Computer Vision that aims to generate answers for the given questions for 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., for developing the VQA model and analysis. Further, to justify the predicted outcome, explanations are required to answer how and why questions. For explanations, Explainable AI (XAI) techniques like Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP) are used.

1.1 MOTIVATION

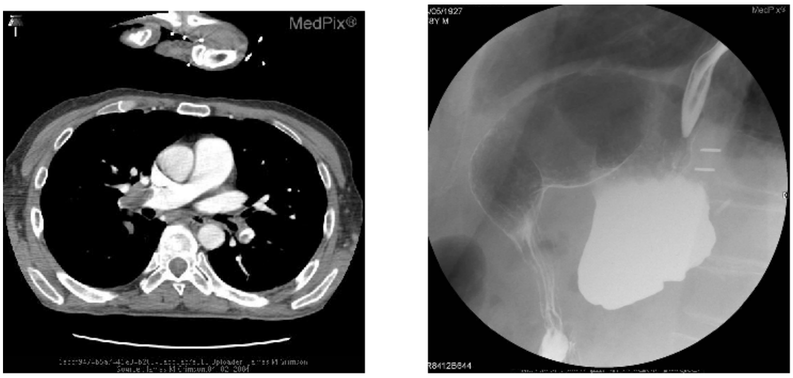
The medical domain is one, where new viruses and new diseases are emerging and needs analysis of patients' conditions. Applying Artificial Intelligence techniques in the Medical field is more effective, where deep analysis of problems can be performed with the help of different AI techniques or algorithms. 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 [27]. 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 the required information. There are Visual Question Answering models for the medical domain that could generate answers for questions, but they lack justification for the predicted outcome [1–4, 7, 9, 10, 12, 17, 20, 21, 25]. To overcome this limitation, the proposed model uses an XAI technique to justify the outcome of the VQA model.

1.2 PROBLEM STATEMENT

This project aims 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 and is validated using XAI techniques. 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 question-answer pairs from ImageCLEF 2019 VQA-Med Dataset (Image IDs: synpic191614, synpic28495)

Objectives

- To collect and analyze the CLEF2019 Medical 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 ORGANISATION OF REPORT

The report is organized as follows: a 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, algorithms and implementation is explained. Chapter 4 illustrates the results of the implementation and performance analysis. Chapter 5 explains the social impact and sustainability of the proposed project. The conclusion and future work are summarized in Chapter 6.

CHAPTER 2

LITERATURE SURVEY

This section discusses various research papers on Visual Question Answering (VQA) for medical images with its techniques and limitations. Also, different Explainable AI (XAI) techniques used with Deep Learning algorithms for the medical domain are discussed.

2.1 VISUAL QUESTION ANSWERING

Visual Question Answering (VQA) task involves generating answers for the given question with the input image. VQA can be applied to several types of images like natural images, cartoon images and medical images. Among the three types of images, VQA for medical images plays a vital role.

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 four questions of each category. The following paragraphs provide a brief overview of the works carried out using ImageCLEF 2019 dataset.

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

(CNN) models for each category of the questions 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 which consists of 2016 questions & image in training and 321 questions & images in validation (used as test dataset). A CNN model is built and trained for predicting the modality of the images. This resulted in an 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 fed to an LSTM model to predict the answer using a classification approach which resulted in an accuracy of 46.2% and a BLEU Score of 48.6.

An encoder-decoder model is proposed by Imane Allaouzi et al., [12] 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 in 55.6% accuracy and a 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 in 57% accuracy and a BLEU Score of 59.1.

Yangyang Zhou et. al., [25] 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 which 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 in 60.6% accuracy and a 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) 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., [12]. This resulted in an accuracy of 63.8%.

Abhishek Thanki et al., [2] used an 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 in an accuracy of 15.5% and a BLEU Score of 45.5.

Lei Shi et al., [21] 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 are used to generate answers. For abnormality, a multi-label classification model is used where each of the output labels corresponds to the part of the answer. This resulted in 56.6% accuracy and a 59.3 BLEU Score.

Attention mechanisms are used with VQA tasks to extract the image feature based on the question. An attention module is used by Shengyan Liu et al., [20] and Minh H. Vu et. al., [17] to get the global image features from the features extracted using pre-trained Xception model and ResNet-152 respectively. The attention module accepts both the image and question features and outputs the global image features. Shengyan Liu et al., [20] then fed the global image features and the question features to a softmax layer for answer prediction which resulted in 21% accuracy and a 39.3 BLEU Score. Minh H. Vu et. al., [17] used bi-linear transformations on the global image feature and question features. The resultant feature is used to predict the answer. This approach with attention module resulted in 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 in an accuracy of 64% and a BLEU Score of 65.9.

The research of VQA, explained in the above paragraphs for ImageCLEF 2019 dataset is summarized in Table 2.1.

TABLE 2.1: Literature survey - VQA for ImageCLEF 2019 dataset

Paper Title	**Methodology**	**Limitations / Inference**
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. Image Feature Extraction: VGGNet16 Answer Generation: Ensemble of Classification models Analysis: Accuracy-60.8, BLEU Score-63.4	The type of the questions in real time is unknown and hence choosing the model for a corresponding question is another 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 Analysis: Accuracy-83.8	Only Modality based questions are considered and within that only major categories of modalities are predicted. The training dataset includes 2016 questions & images and the validation dataset includes 321 questions & images which are used for testing.
Tlemcen University at ImageCLEF 2019 Visual Question Answering Task [7]	Image Feature Extraction: VGGNet16 Question Feature Extraction: LSTM Answer Generation: LSTM Analysis: Accuracy-46.2, BLEU-48.6	Features can be extracted using pre-trained models from both images and questions. The features can be used to generate answers.

An Encoder-Decoder model for visual question answering in the medical domain [12]	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	Features extracted from images and questions can be concatenated and then a fully connected neural network can be used to generate answers from the concatenated features.
JUST at ImageCLEF 2019 Visual Question Answering in the Medical Domain [4]	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	The type of the questions in real time is unknown and hence choosing the model for a corresponding question is another task.
TUA1 at ImageCLEF 2019 VQA-Med: A classification and generation model based on transfer learning [25]	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	The type of the questions in real time is unknown and hence choosing the model for a corresponding question is another task.

MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain [9]	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]	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 [21]	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	In the case of multi-label classification, the predicted labels are in random order and techniques to order the labels need to be implemented.

An Xception-GRU Model for Visual Question Answering in the Medical Domain [20]	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	An efficient attention module can be used for extracting the best features from the image.
Ensemble of Streamlined Bilinear Visual Question Answering Models for the ImageCLEF 2019 Challenge in the Medical Domain [17]	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	An image pre-processing pipeline is used to pre-process the image and remove unwanted information from the image.
CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering [10]	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 questions such as Modality, Plane, Organ and Abnormality. But in reality, the exact type of question is not known.

2.2 EXPLAINABLE AI (XAI)

Explainable AI refers to techniques and tools that help to analyze and interpret the predictions made by Machine Learning (ML) / Deep Learning (DL) models [23]. These tools analyze the model and find the reasons behind the predicted outcome of the model [11].

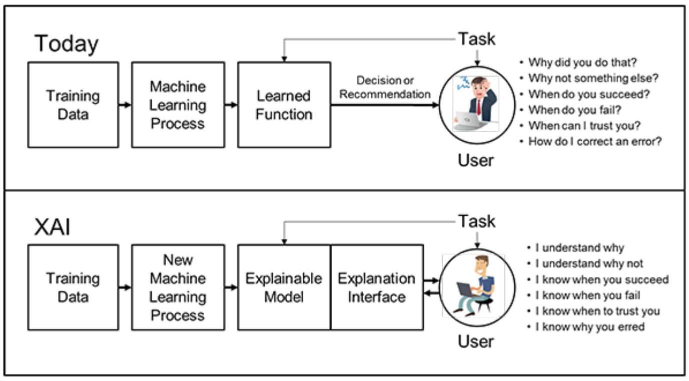


FIGURE 2.1: Traditional-AI vs XAI

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 and trust in the AI systems are increases [23].

- **Result tracing:** XAI techniques are used to trace the features of the input that affect the outcome of the AI system [23].

- **Model improvement:** In general, AI systems learn from the 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 the correctness of the decision can be validated [23].

2.2.1 XAI Techniques

Many XAI techniques have been widely used for different types of architecture. Some of the XAI techniques are:

- Local Interpretable Model-Agnostic Explanations (LIME)
- SHapley Additive exPlanations (SHAP)
- Contextual Importance and Utility (CIU)
- Class Activation Mapping (CAM)
- Gradient-weighted Class Activation Mapping (Grad-CAM)
- Anchors
- Integrated Gradients

The following paragraphs provide a brief overview of the works involving various XAI techniques for interpreting the outcome of models in medical doamin.

Knapič S et al., [14] 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 LIME, SHAP and CIU are used 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 CAM, Grad-CAM, Anchors and Integrated Gradients which are used to generate explanations on the predicted outcome of various models.

CAM helps 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., [13] 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 justify the predicted outcome.

Similarly, Yu-Huan Wu et. al., [26] used a ResNet model for classifying CT scan images of COVID-19. The explainability is provided using CAM.

Ramprasaath R. Selvaraju et al., [18] 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 [22]. The results of each of the above tools are combined to give explanations.

From the study and analysis, it is inferred that, LIME and SHAP are suitable for analyzing VQA Model.

LIME

Local Interpretable Model-agnostic Explanations (LIME) [15] 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 may be any simple or interpretable model such as linear models or decision trees or falling rule lists (if-then rules). As every $g \in G$ may not be simple and interpretable, the complexity $\Omega(g)$ is taken into account while finding the explanations. The complexity $\Omega(g)$ can be depth of the tree (for decision trees) or the number of non-zero weights (for linear models).

Let f denotes the model to be interpreted and $f(x)$ is the probability that x belongs to a particular 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 Equation 2.1

$$\xi(x) = \underset{g \in G}{\arg\min}\, \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{2.1}$$

SHAP

SHapley Additive exPlanations (SHAP) [19] 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) influences 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 Shapley Values [28]. The Shapley values are computed as given by Equation 2.2.

$$\phi_i(v, F) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [v(S \cup \{i\}) - v(S)] \tag{2.2}$$

where S is all the subsets in F which is the set of players (features) and v is the outcome. The Equation 2.2 calculates the weighted marginal contribution of i. The marginal contribution of i refers to the change in the outcome of the function v such that i is present.

SHAP aims 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 [29]. SHAP calculates the Shapley Values wich is based on Game Theory. A feature value can act alone or as a group to arrive at an outcome. The Shapley values specify 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 Equation 2.3.

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j^{'} \tag{2.3}$$

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.

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

TABLE 2.2: Literature survey - XAI techniques

Paper Title	**Dataset**	**Model(s) used**	**XAI technique used**
Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain [14]	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 [13]	Lung Ultrasound (LUS)	VGGNet	Class Activation Mapping (CAM)

JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation [26]	COVID-19 Classification and Segmentation (COVID-CS) dataset	ResNet	Class Activation Mapping (CAM)
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization [18]	-	-	Gradient-weighted Class Activation Mapping (Grad-CAM)
LISA : Enhance the explainability of medical images unifying current XAI techniques [22]	COVID-19 Dataset	CNN (Pre-trained)	LIME, SHAP, Anchors, Integrated Gradients

2.3 RESEARCH GAP

(i) Some of the existing VQA research works [3, 4, 10, 25] uses different models for a different type of questions, but the question type, in general, is not known.

(ii) The existing VQA models generate answers but do not provide explanations for the outcome [9, 12].

(iii) The performance of VQA models has been analyzed using Quantitative Metrics like Accuracy, BLEU Score and WBSS, but had not been analyzed qualitatively using XAI techniques.

2.4 RESEARCH OBJECTIVES

- To address the Research Gap (i), features from questions are extracted using the BERT tokenizer and concatenated to the image features, extracted using VGGNet. The concatenated features are used to generate answers using BERT.

- To address the Research Gaps (ii) & (iii), XAI techniques LIME and SHAP are used for generating explanations on the outcome of the VQA model and analyze performance of the model qualitatively. The explanations not only provide an analysis of the prediction but also justify the generated answers.

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 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.

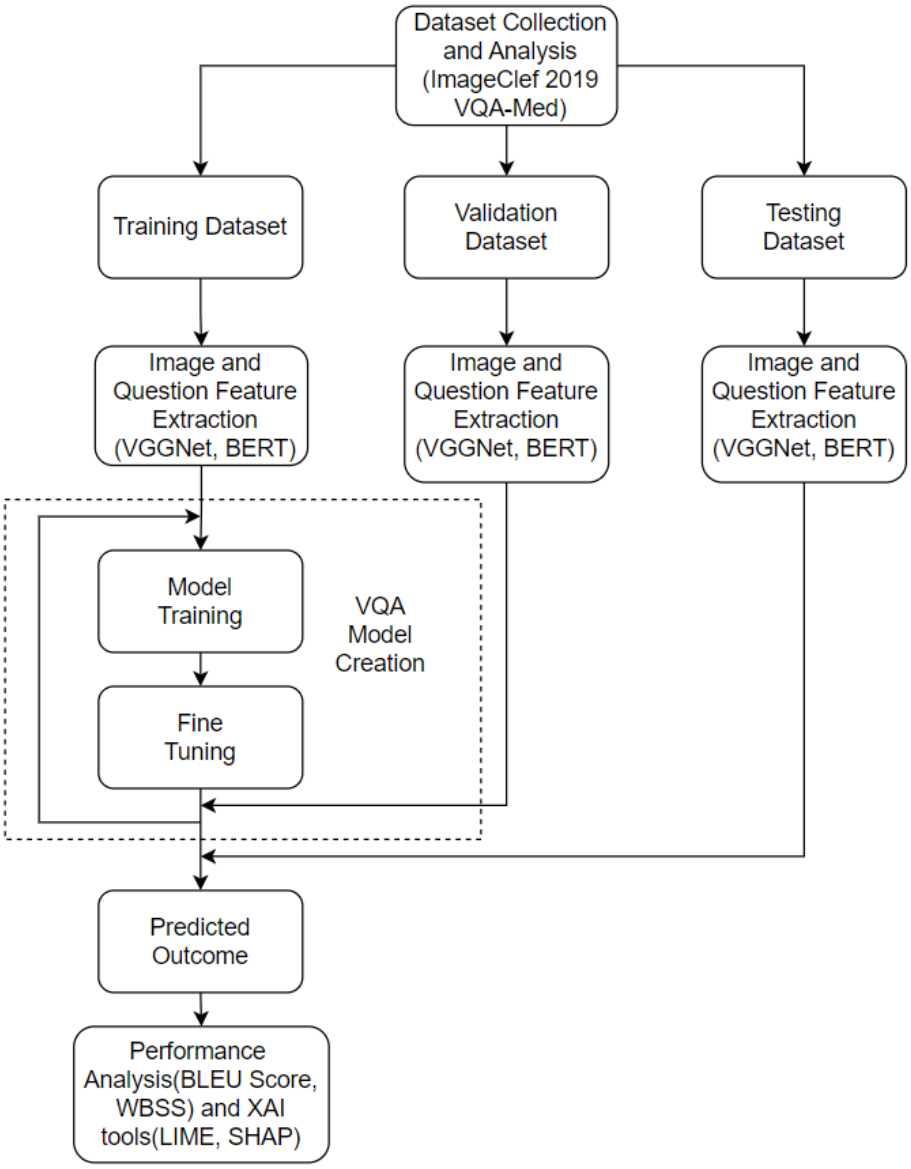

Dataset Collection and Analysis (ImageClef 2019 VQA-Med)
Training Dataset
Validation Dataset
Testing Dataset
Image and Question Feature Extraction (VGGNet, BERT)
Image and Question Feature Extraction (VGGNet, BERT)
Image and Question Feature Extraction (VGGNet, BERT)
Model Training
VQA Model Creation
Fine Tuning
Predicted Outcome
Performance Analysis(BLEU Score, WBSS) and XAI tools(LIME, SHAP)

FIGURE 3.1: Proposed 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 in 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 are discussed in the following subsections.

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 modified architecture of VGGNet is shown in Figure 3.2.

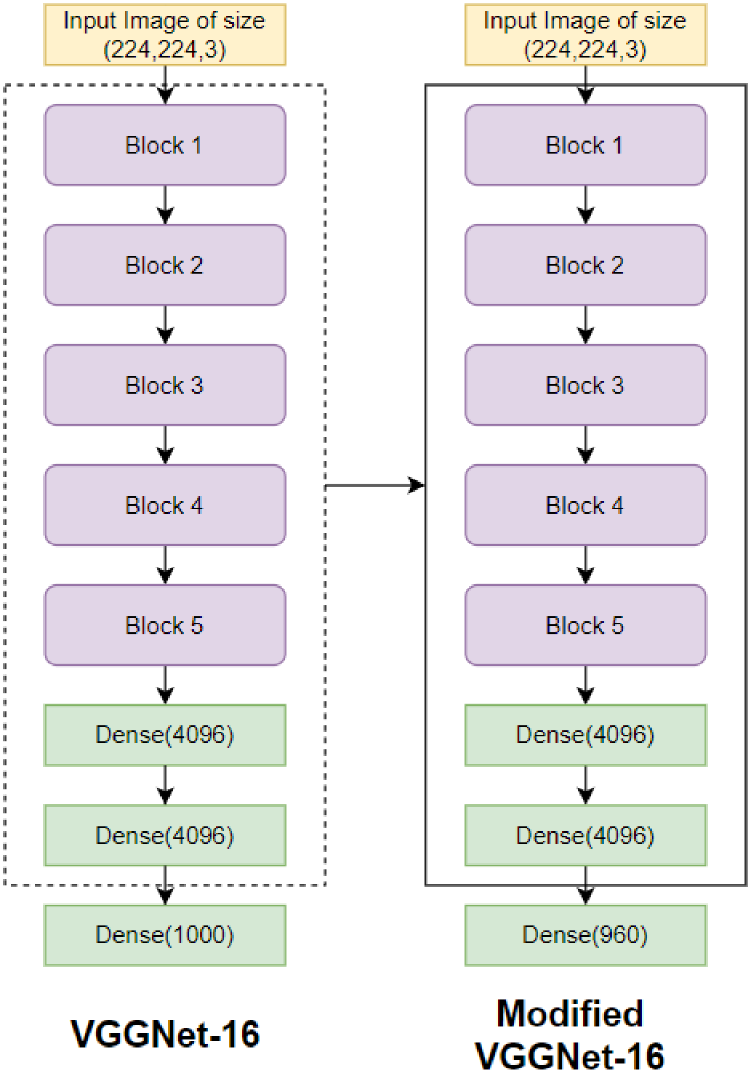
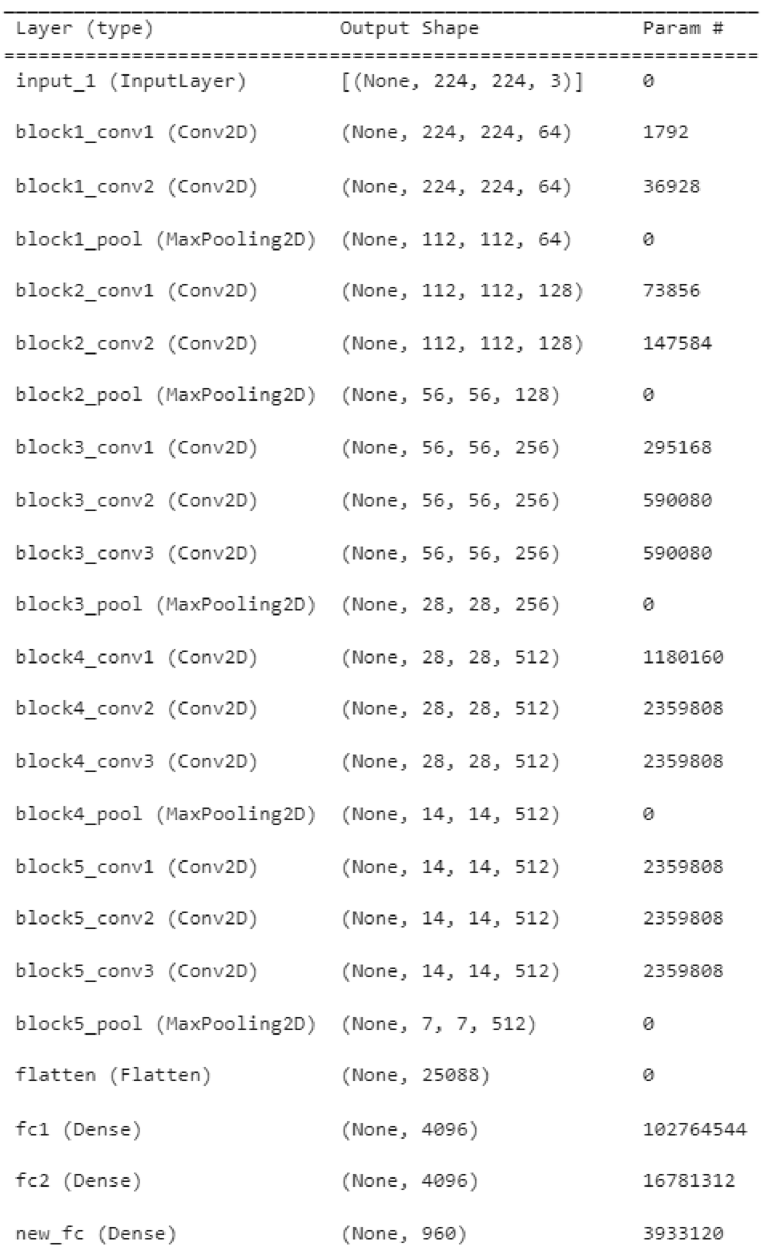


FIGURE 3.2: Modified VGGNet architecture

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. The image feature encoding is depicted in 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. Hence the question needs to be pre-processed by removing these special characters & punctuation and also the text is converted to lower casedr. 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)
    $Lower\_Case\_Qn \leftarrow Question.lower()$
    $Preprocessed\_Question \leftarrow Lower\_Case\_Qn.replace("[^\wedge a-z0-9]", " ")$
    return $Preprocessed\_Question$
end function
```

3.2.4 Question Feature Extraction

The question features are the set of tokens generated through 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 into tokens (Question Features). The Question Feature Extraction Process is explained in Algorithm 4.

Algorithm 4 Question Feature Encoding

```
Input : $Question$                                      ▷ Input Question for Feature Extraction
Output : $Question\_Encoding$
$Tokenizer \leftarrow BERT\_Tokenizer(VocabFilePath)$   ▷ Question Tokens
function GETTOKENIZEDQUESTION(Question)
    $Preprocessed\_Question \leftarrow textPreProcess(Question)$
    $Question\_Encoding \leftarrow Tokenizer(Preprocessed\_Question)$
    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 a model for answer generation 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 which takes 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 answer tokens are masked and the model is trained to predict the masked tokens. 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.

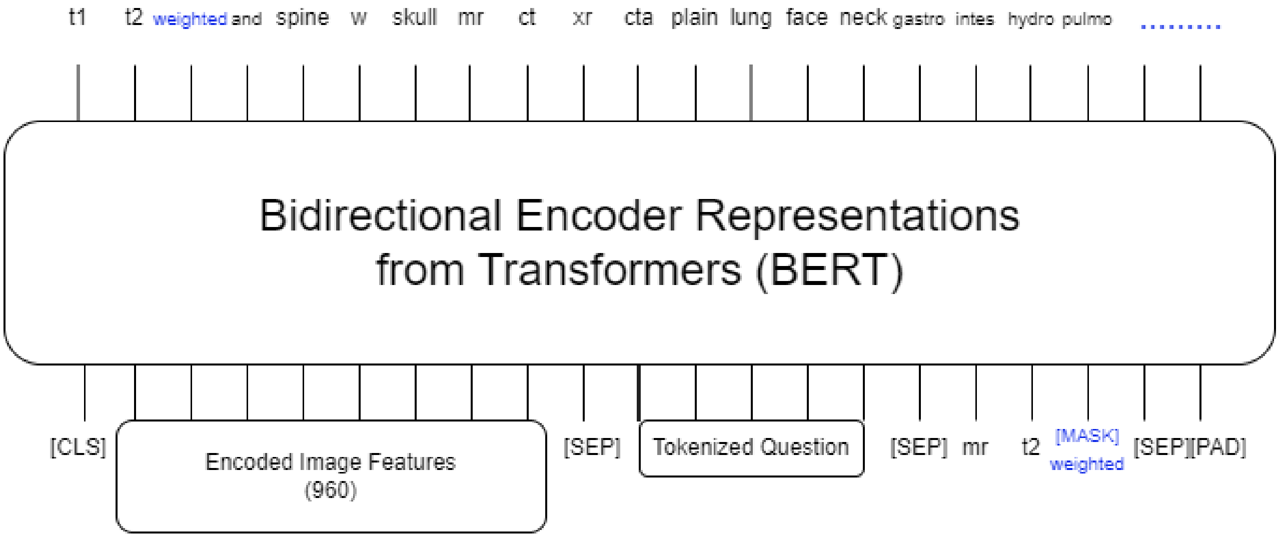


FIGURE 3.4: BERT model predicting the masked word

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

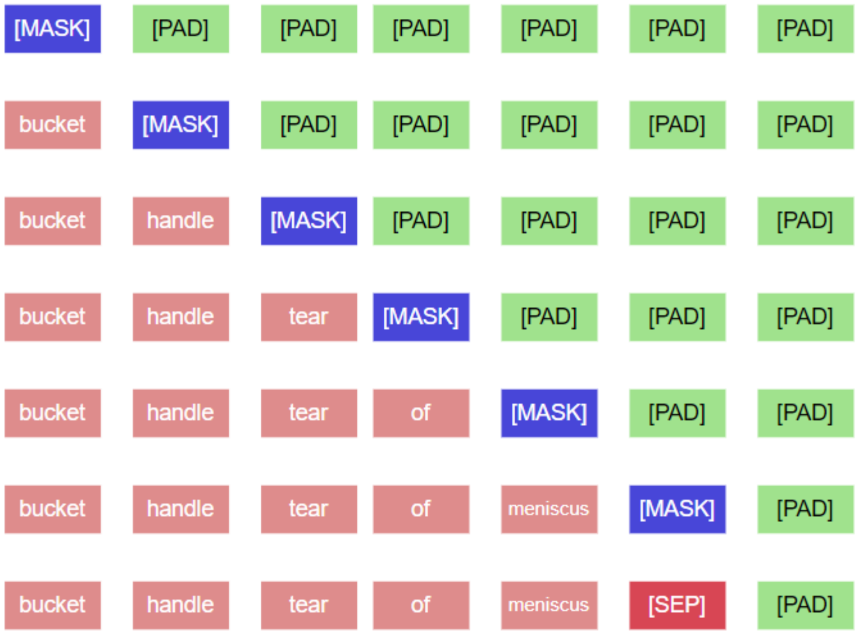
[CLS] ENCODED-IMAGE-FEATURES [SEP] QUESTION-FEATURE [SEP] [MASK]

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 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 Figure 3.5



[MASK]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]
bucket	[MASK]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]
bucket	handle	[MASK]	[PAD]	[PAD]	[PAD]	[PAD]
bucket	handle	tear	[MASK]	[PAD]	[PAD]	[PAD]
bucket	handle	tear	of	[MASK]	[PAD]	[PAD]
bucket	handle	tear	of	meniscus	[MASK]	[PAD]
bucket	handle	tear	of	meniscus	[SEP]	[PAD]

FIGURE 3.5: Answer generation for a sample with ID: synpic58267

3.4 EXPLAINABLE AI (XAI) TECHNIQUES

Explainable AI (XAI) techniques helps to interpret the predictions of ML/DL models. XAI analysis the output with respect to the input features that contributed to the prediction. There are many XAI techniques like LIME, SHAP, Anchor, Contextual Importance and Utility (CIU), Gradient-weighted Class Activation Mapping (Grad-CAM) and etc. In this project, LIME and SHAP are used for analyzing the outcome with explanations.

3.4.1 XAI - LIME

Local Interpretable Model-agnostic Explanations (LIME) generates explanations by building a local interpretable model on the perturbed dataset. A perturbed dataset is a dataset that is created by modifying the incoming input instance. The instances of the perturbed dataset are fed to the model to be interpreted. The output of the model is analyzed to examine if the modifications on the actual input have been reflected in the output. This analysis is used to build the local model, which can be a linear model, a decision tree, or a falling rule list (if-then rule list). The actual input instance is fed to this local model, and the explanations are generated by tracing the flow of the input.

In this project, LIME Image Explainer is used for generating explanations. LIME image explainer accepts a prediction function that accepts the input for the model, an instance to explain, inpaint color (shades of grey), i.e., hide color, and the 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, this is an interpretable model. This model is interpreted and analyzed to find the segments of the image that affect the output, thereby predicting the same outcome as that of the actual model to be interpreted.

3.4.2 XAI - SHAP

SHapley Additive exPlanations (SHAP) generate explanations by computing Shapley values for the features of the input instance. These Shapely values represent the contributions of the features to the output. Shap provides various types of explainers, like SHAP Deep Explainer, SHAP Kernel Explainer, SHAP Tree Explainer, SHAP Gradient Explainer, SHAP Sampling Explainer, SHAP Partition Explainer, SHAP Permutation Explainer, etc. For tasks involving images like image classification and image captioning, SHAP Partition Explainer is used to interpret the outcome of the model. For visualizing the explanations, SHAP provides various formats like SHAP bar plot, waterfall plot, scatter plot, heatmap, image, and partial dependence plot.

In this project, SHAP Partition Explainer is used for generating explanations on the outcome of the developed VQA model. 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 to mask the regions of the image and find the impact of masking a region. The impact is computed in terms of Shapley Values. The Shapley values of each feature in the input are calculated from the Shapley values of the combinations of features by finding the global contribution, and the computed Shapley values are visualized using SHAP's Image Plot.

3.5 QUANTITATIVE METRICS FOR VQA MODEL

The performance of Visual Question Answering Models can be analyzed using various quantitative metrics such as Accuracy, Bilingual Evaluation Understudy (BLEU) Score and Word-based Semantic Similarity (WBSS). Accuracy is a measure of how accurately the generated answer matches the actual answer. The perfect match gives a score of 1, otherwise 0. Accuracy is calculated as shown in Equation 3.1.

$$Accuracy = \frac{No.\ of\ correctly\ generated\ answers}{Total\ no.\ of\ samples} \tag{3.1}$$

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 many 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 matches 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 are calculated as given by Equation 3.2

$$PrecisionOneGram = \frac{No.\ of\ CorrectlyPredictedOneGrams}{No.\ of\ TotalPredictedOneGrams} \tag{3.2}$$

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 Equation 3.3.

$$BrevityPenalty = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c <= r \end{cases} \tag{3.3}$$

Finally to calculate BLEU Score, the Brevity Penalty is multiplied with Precision 1-gram value as given by Equation 3.4

$$BLEU \;=\; BrevityPenalty\;.\;PrecisionOneGram \tag{3.4}$$

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[16] is calculated as depicted in Equation 3.5.

$$WUPS(A,T) = \frac{1}{N} \times \sum_{i=1}^{N} \times min\left\{ \prod_{a \in A^i} max_{t \in T^i} WUP(a,t), \prod_{t \in T^i} max_{a \in A^i} WUP(a,t) \right\} \times 100 \tag{3.5}$$

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 DATASET ANALYSIS

The ImageCLEF 2019 VQA-Med dataset is used for developing the VQA model. 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 a total of 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 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	1553
Validation	500	Modality	500	35
		Plane	500	15
		Organ System	500	10
		Abnormality	500	413
		Total	2000	473
Testing	500	**All**	500	166

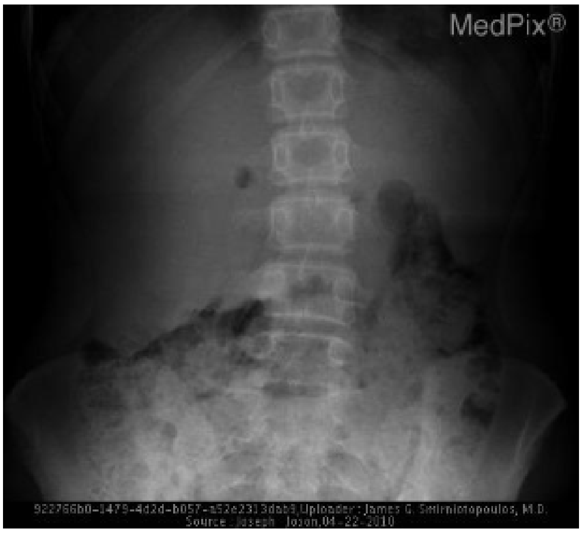
TABLE 4.1: Dataset analysis

The dataset has three types of samples.

(i) An image with single question-answer pair (shown in Figure 4.1).

(ii) An image with multiple question-answer pairs (shown in Figure 4.2).

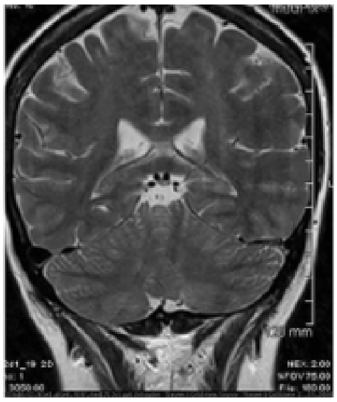
(iii) Multiple images with same question-answer pair (shown in Figure 4.3).


MedPix®
922766b0-1479-4d2d-b057-a52e2313dab9,Uploader : James G. Smirniotopoulos, M.D.
Source : Joseph Jozon,04-22-2010

which organ system is shown in the x-ray?

spine and contents

FIGURE 4.1: Single image with single question (Image ID: synpic52980)



- **is this a t1 weighted, t2 weighted, or flair image?**
  - T2
- **what imaging plane is depicted here?**
  - Coronal
- **what organ system is shown in the image?**
  - skull and contents
- **what is abnormal in the mri?**
  - colloid (neuroepithelial) cyst of the third ventricle

FIGURE 4.2: Single image with multiple questions (Image ID: synpic16994)

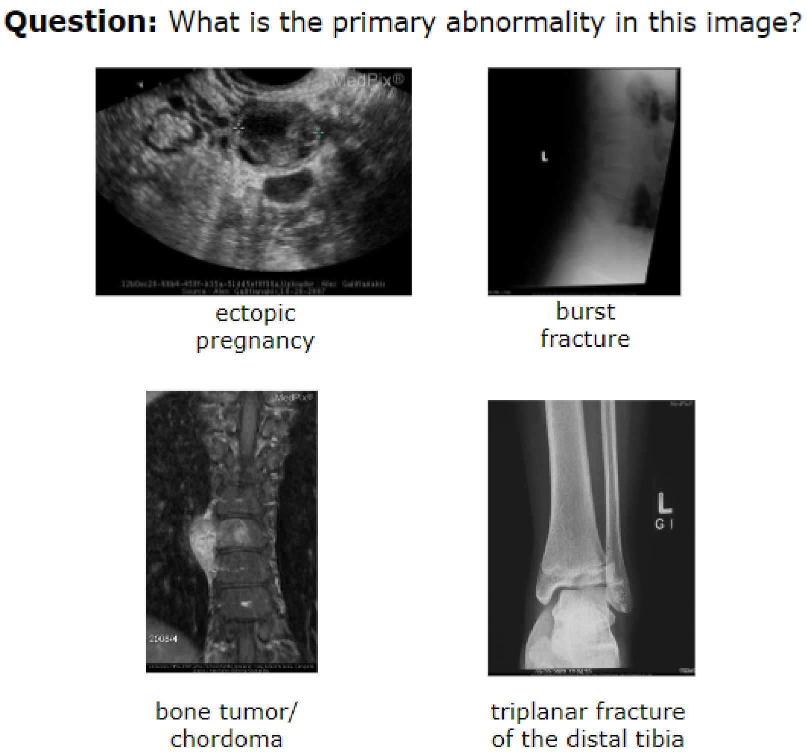

Question: What is the primary abnormality in this image?
ectopic pregnancy
burst fracture
bone tumor/ chordoma
triplanar fracture of the distal tibia

FIGURE 4.3: Multiple images with same question (Image IDs: synpic38930, synpic52143, synpic20934, synpic19141)

4.2 ECOSYSTEM

The hardware and software specifications for the proposed project are as follows:

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

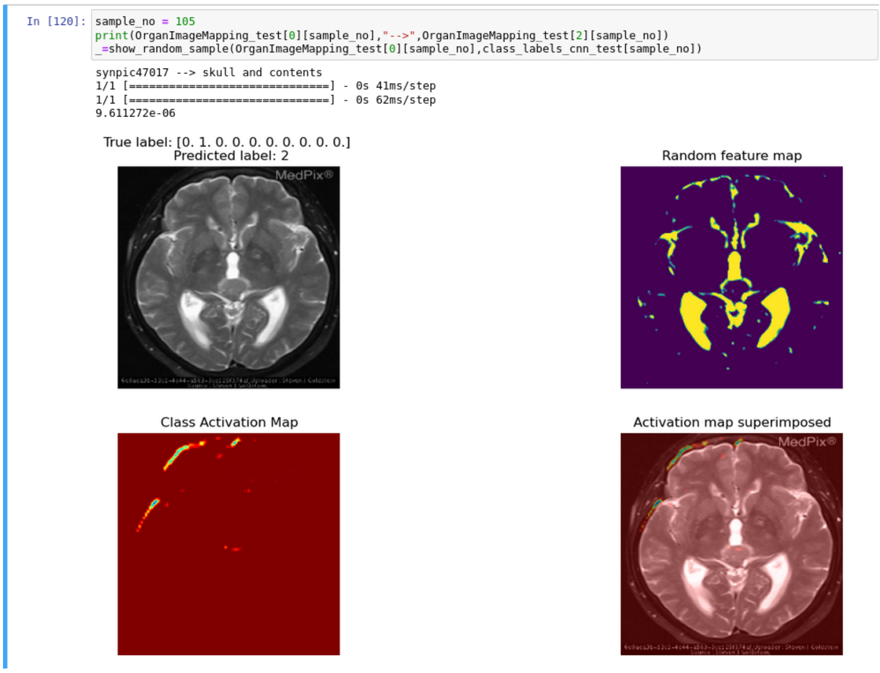
4.3 RESULT OF FEATURE EXTRACTION

The images and questions are pre-processed and the features are extracted using VGGNet and Bidirectional Encoder Representations from Transformers (BERT) respectively.

4.3.1 Result of Image Feature Extraction

The image features are extracted and are encoded as explained in Algorithm 2 in the section 3.2 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.



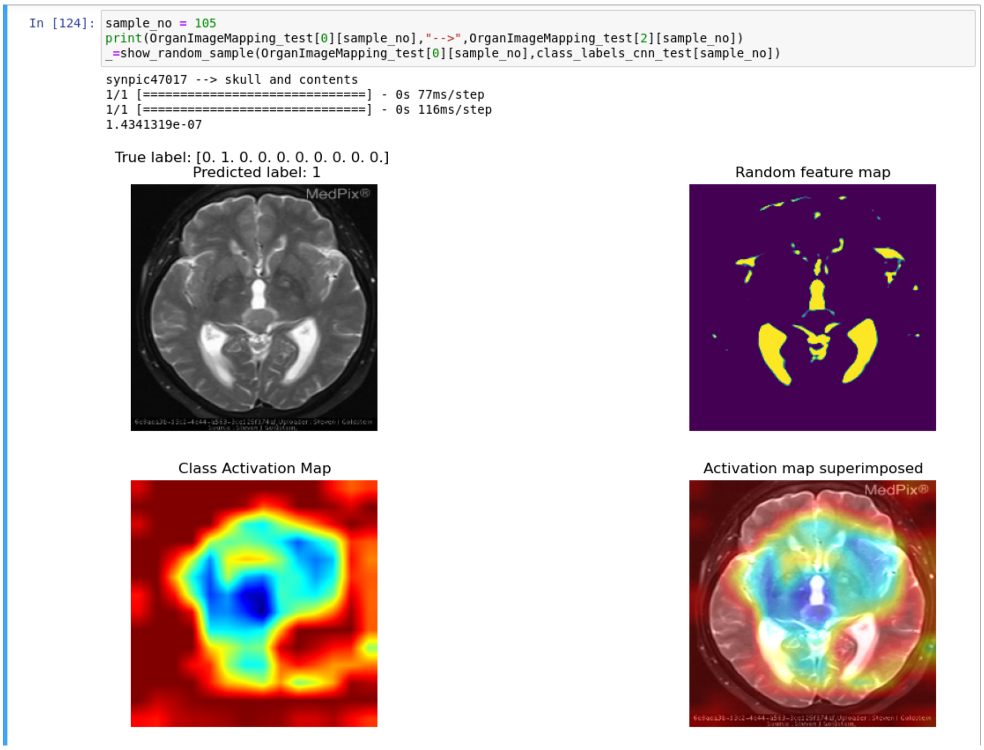
```
In [120]: sample_no = 105
          print(OrganImageMapping_test[0][sample_no],"-->",OrganImageMapping_test[2][sample_no])
          _=show_random_sample(OrganImageMapping_test[0][sample_no],class_labels_cnn_test[sample_no])

          synpic47017 --> skull and contents
          1/1 [==============================] - 0s 41ms/step
          1/1 [==============================] - 0s 62ms/step
          9.611272e-06
```

True label: [0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
Predicted label: 2
MedPix®
Random feature map
Class Activation Map
Activation map superimposed
MedPix®

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.



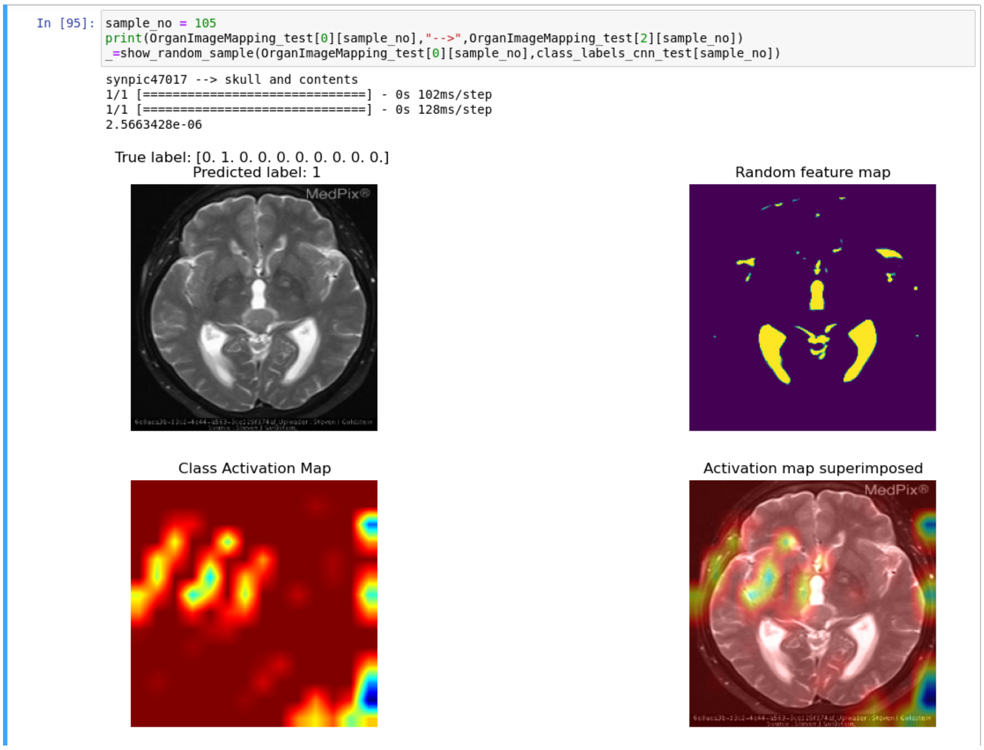
```
In [124]: sample_no = 105
          print(OrganImageMapping_test[0][sample_no],"-->",OrganImageMapping_test[2][sample_no])
          _=show_random_sample(OrganImageMapping_test[0][sample_no],class_labels_cnn_test[sample_no])

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

True label: [0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
Predicted label: 1
MedPix®
Random feature map
Class Activation Map
Activation map superimposed
MedPix®

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.



```
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
```

True label: [0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
Predicted label: 1
MedPix®
Random feature map
Class Activation Map
Activation map superimposed
MedPix®

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.

4.3.2 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].

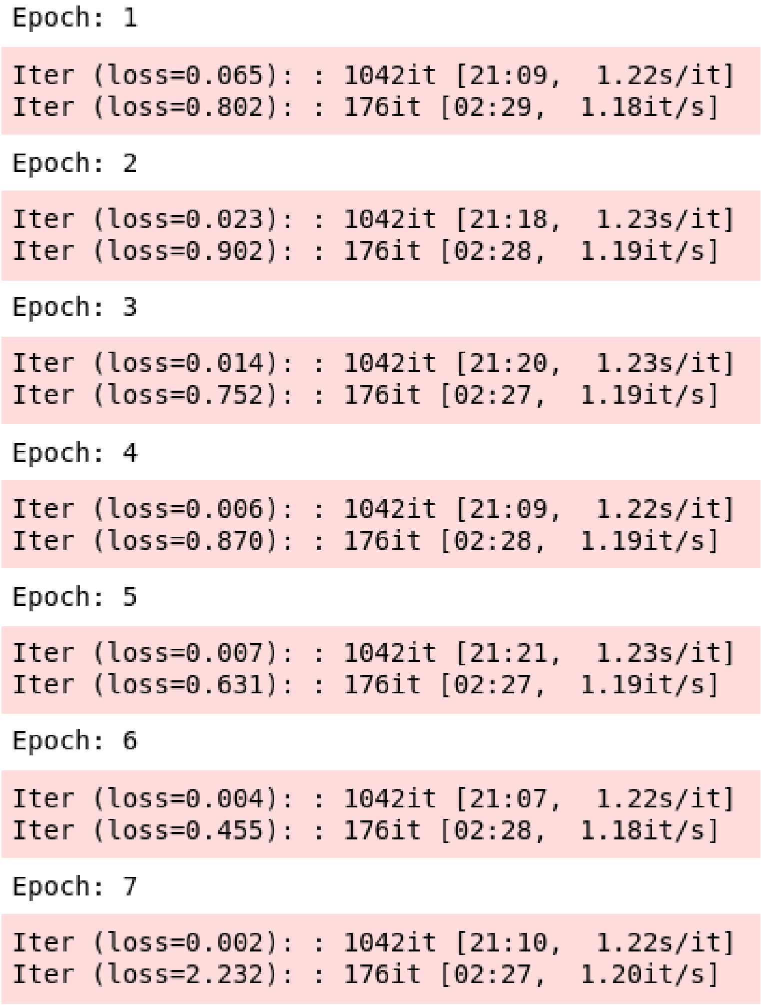
A sample set of tokens and their corresponding token ID is shown in Table 4.2.

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

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

4.4 RESULT OF VQA MODEL

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

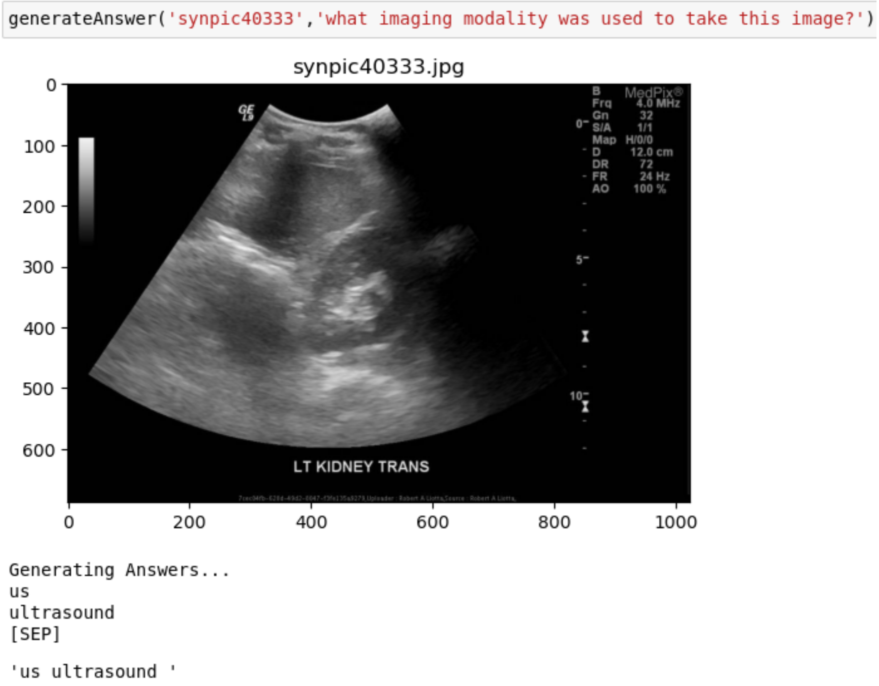


```
Epoch: 1
Iter (loss=0.065): : 1042it [21:09,  1.22s/it]
Iter (loss=0.802): : 176it [02:29,  1.18it/s]
Epoch: 2
Iter (loss=0.023): : 1042it [21:18,  1.23s/it]
Iter (loss=0.902): : 176it [02:28,  1.19it/s]
Epoch: 3
Iter (loss=0.014): : 1042it [21:20,  1.23s/it]
Iter (loss=0.752): : 176it [02:27,  1.19it/s]
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, 4.9 and 4.10 show samples of correct answers generated by the model when queried about Modality, Plane and Organ respectively. Figure 4.11 shows the answer generated by the model when queried about Abnormality. The answer generated is similar to the correct answer (*pancreatic ductal adenocarcinoma*).

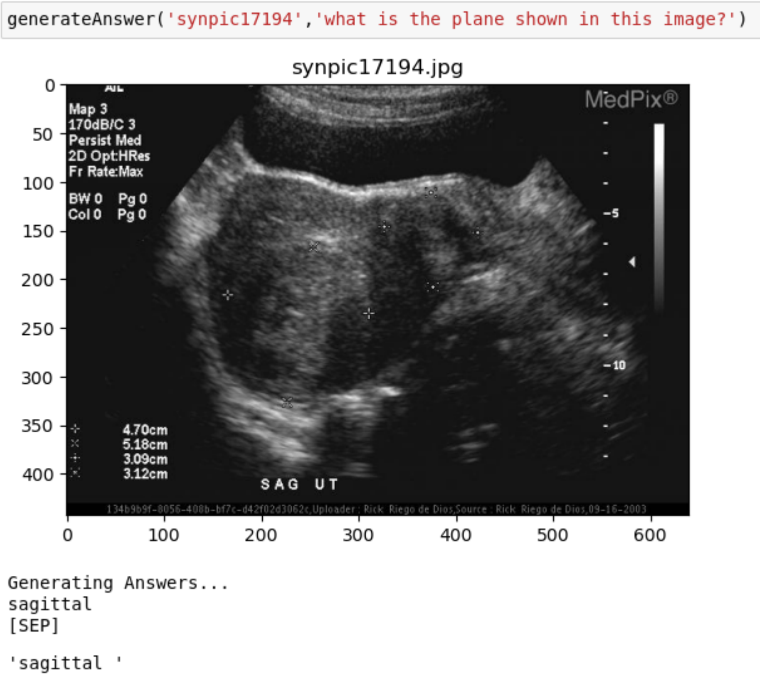


```
generateAnswer('synpic40333','what imaging modality was used to take this image?')
```

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

'us ultrasound '
```

FIGURE 4.8: Answer generation for the image (ID: synpic40333) queried about modality

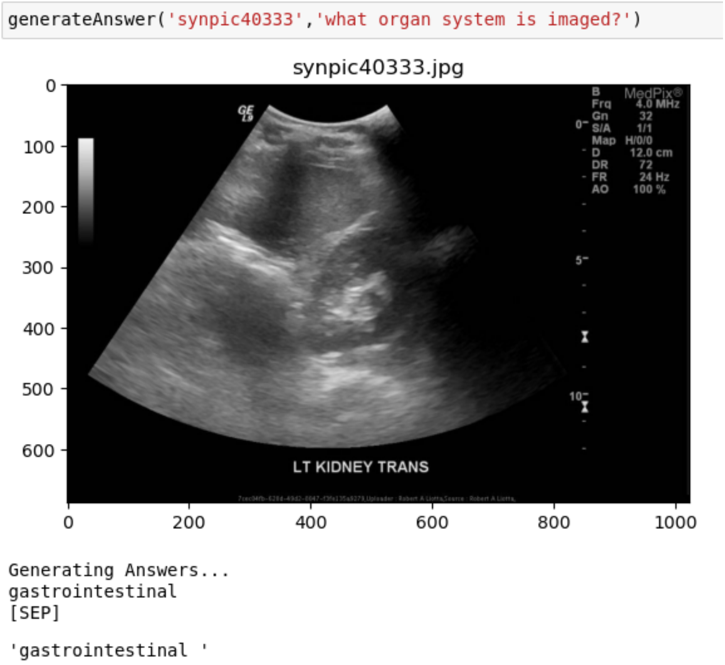


```
generateAnswer('synpic17194','what is the plane shown in this image?')
```

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

'sagittal '
```

FIGURE 4.9: Answer generation for the image (ID: synpic17194) queried about plane

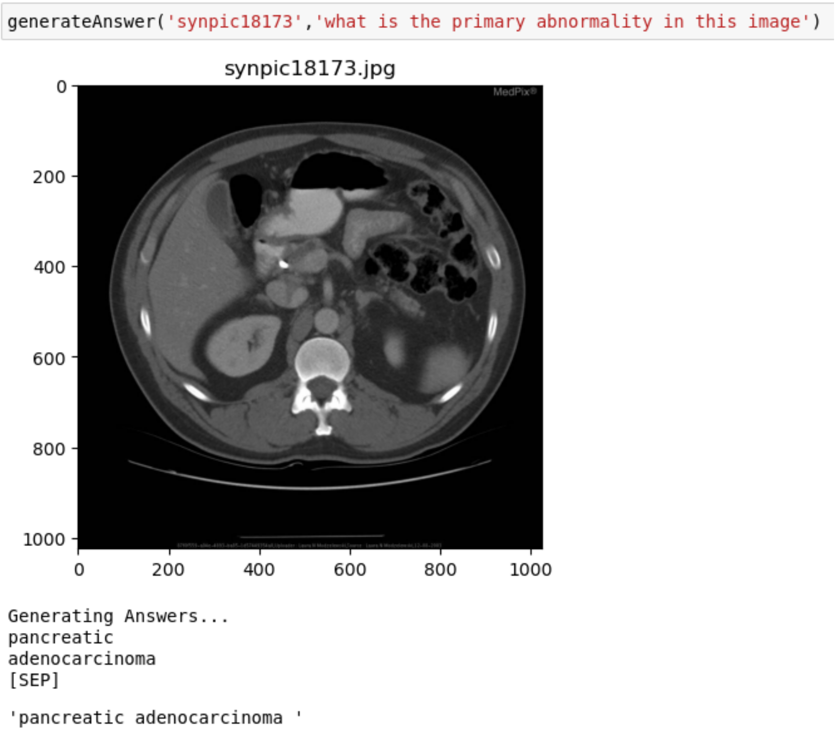


```
generateAnswer('synpic40333','what organ system is imaged?')
```

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

'gastrointestinal '
```

FIGURE 4.10: Answer generation for the image (ID: synpic40333) queried about organ



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

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

'pancreatic adenocarcinoma '
```

FIGURE 4.11: Answer generation for the image (ID: synpic18173) queried about abnormality

4.5 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.

4.5.1 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 five samples (one actual input image + four generated images) is shown in Figure 4.13

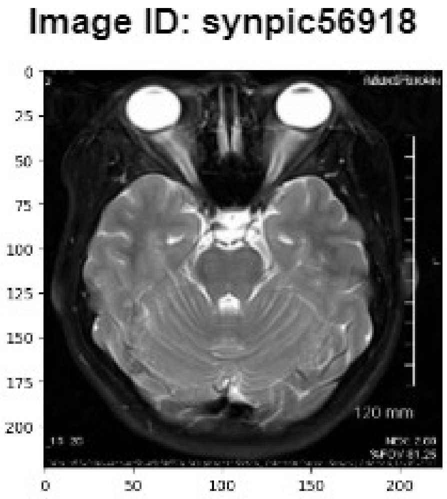


FIGURE 4.12: Sample input for LIME (Image ID: synpic56918)

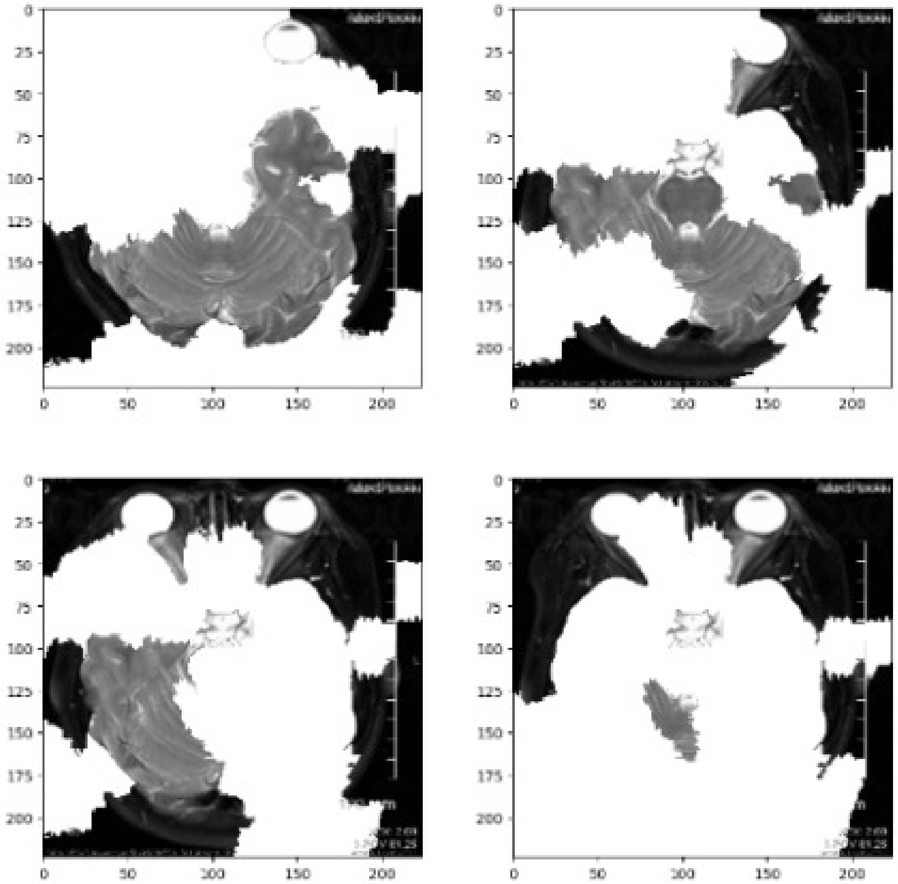


FIGURE 4.13: Perturbed dataset for the sample input (ID: synpic56918)

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

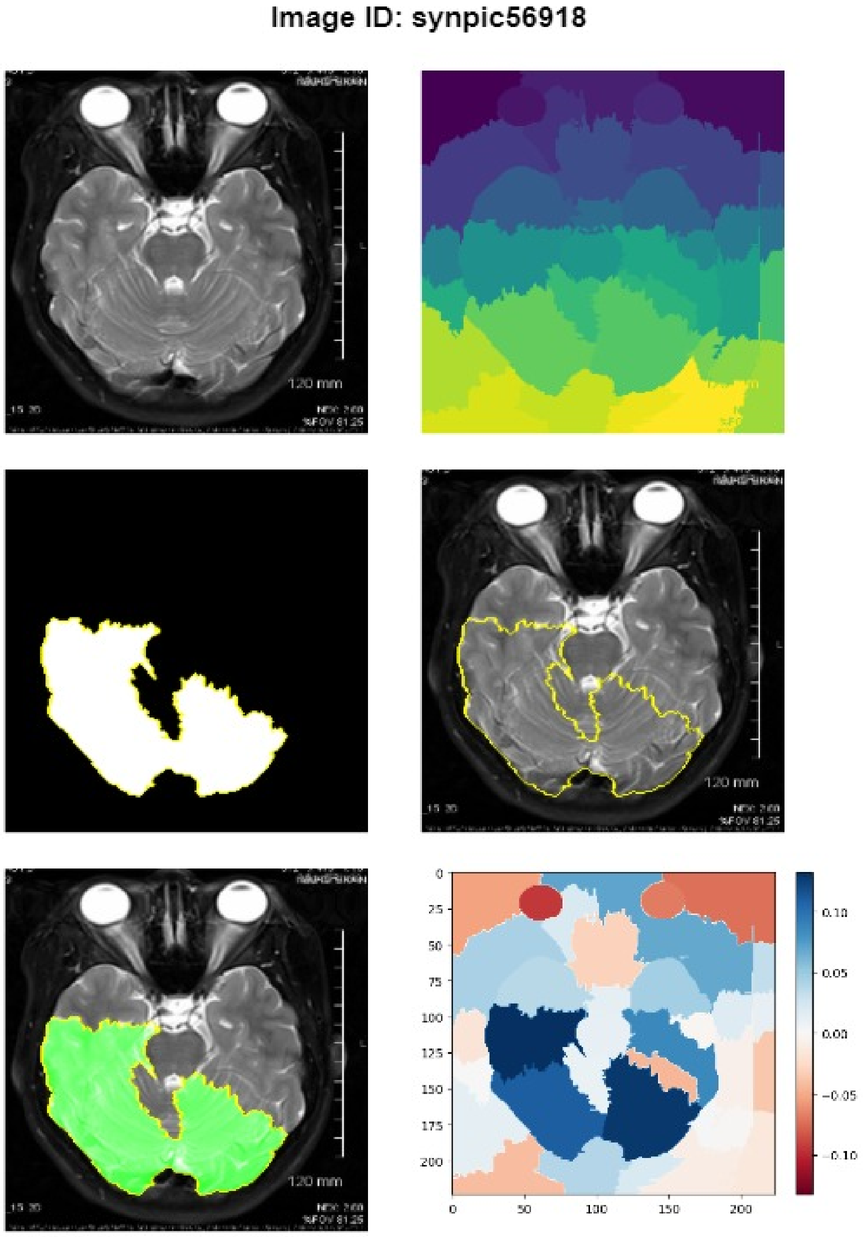


FIGURE 4.14: LIME explanation for a sample image (ID: synpic56918) queried about organ

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.

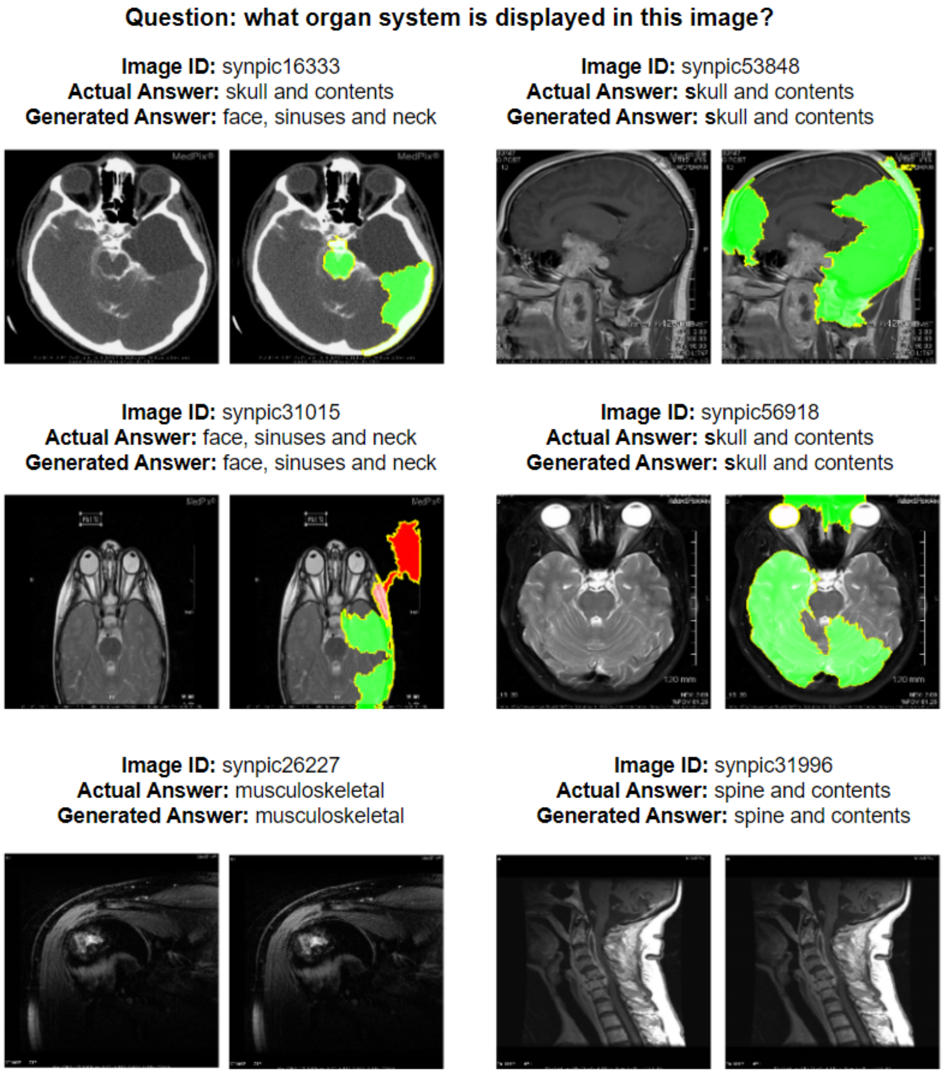

Question: what organ system is displayed in this image?
Image ID: synpic16333
Actual Answer: skull and contents
Generated Answer: face, sinuses and neck
Image ID: synpic53848
Actual Answer: skull and contents
Generated Answer: skull and contents
Image ID: synpic31015
Actual Answer: face, sinuses and neck
Generated Answer: face, sinuses and neck
Image ID: synpic56918
Actual Answer: skull and contents
Generated Answer: skull and contents
Image ID: synpic26227
Actual Answer: musculoskeletal
Generated Answer: musculoskeletal
Image ID: synpic31996
Actual Answer: spine and contents
Generated Answer: spine and contents

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

In Figure 4.15, for the image with ID: synpic31015, there is red inpaint on the output image. The red inpaint indicates a 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 are different.

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

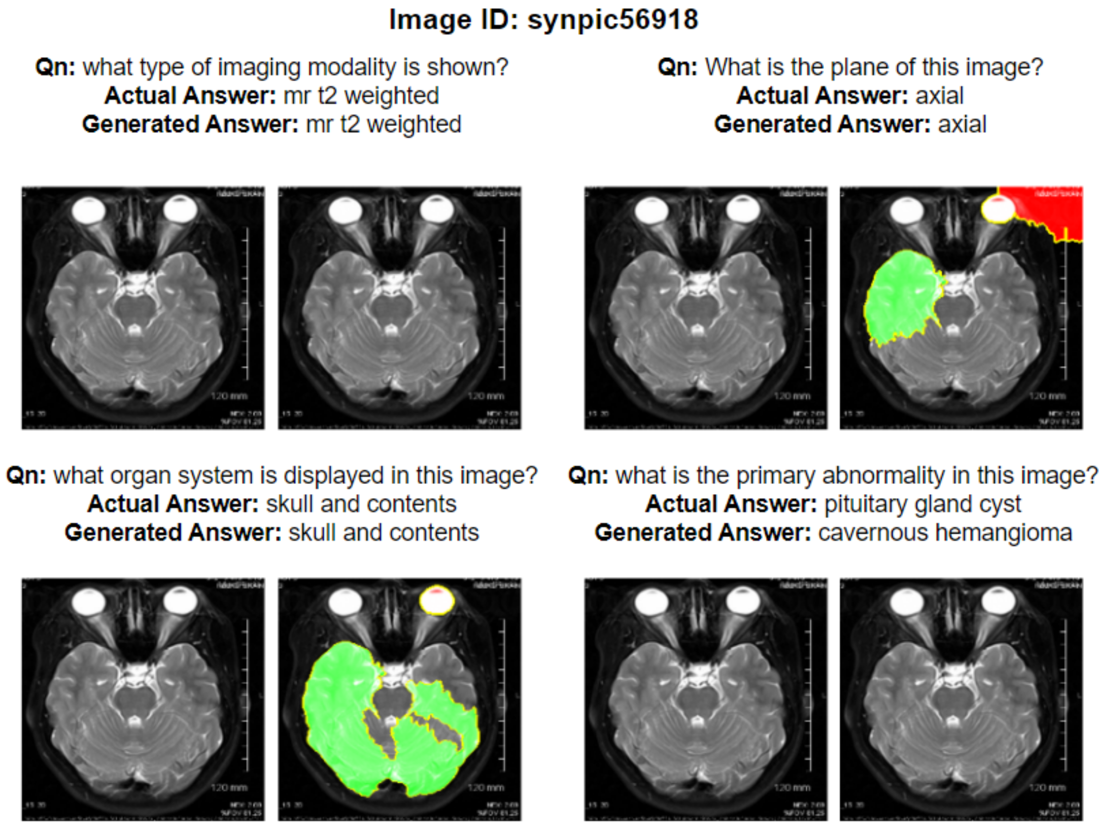


Image ID: synpic56918

Qn: what type of imaging modality is shown?
Actual Answer: mr t2 weighted
Generated Answer: mr t2 weighted

Qn: What is the plane of this image?
Actual Answer: axial
Generated Answer: axial

Qn: what organ system is displayed in this image?
Actual Answer: skull and contents
Generated Answer: skull and contents

Qn: what is the primary abnormality in this image?
Actual Answer: pituitary gland cyst
Generated Answer: cavernous hemangioma

FIGURE 4.16: LIME explanations for a sample image (ID: synpic56918) for four different questions of different categories

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

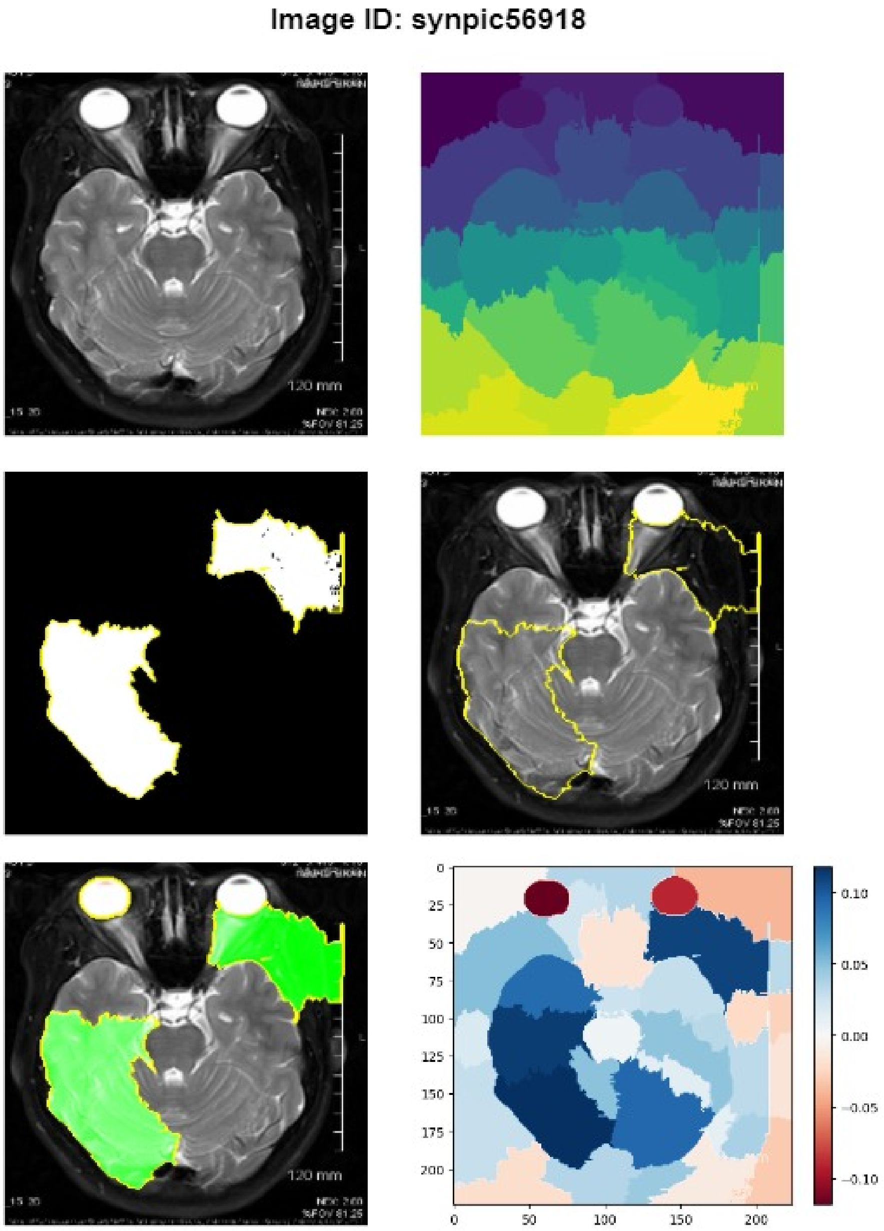


FIGURE 4.17: LIME explanation for a sample image (ID: synpic56918) queried about organ on the second run

From Figures 4.14 and 4.17, it is inferred that the explanations differ in each run. This is because 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.

4.5.2 XAI - SHAP

SHAP calculates Shapely values for each feature in the input instance. The Shapely values measure 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 the output of SHAP for the image with ID: synpic56918 queried about the organ system.

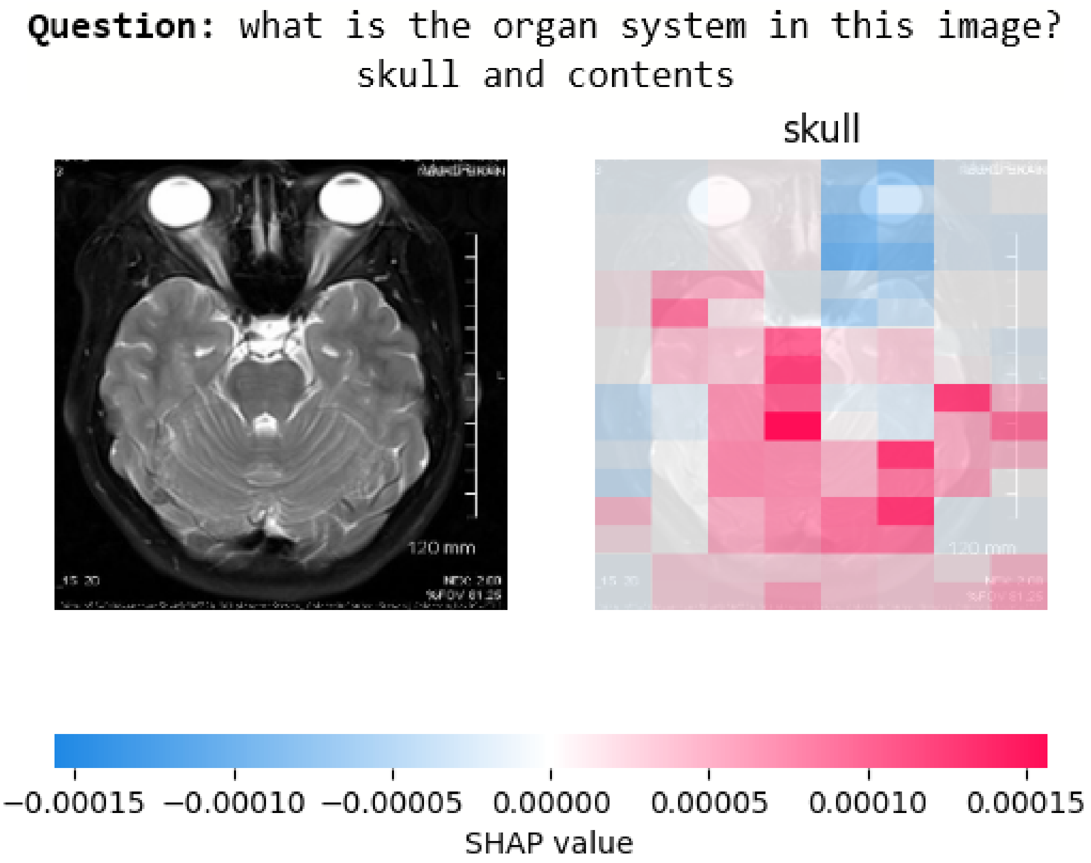


FIGURE 4.18: SHAP explanation for a sample image (ID: synpic56918) queried about organ

The regions highlighted in red color represent the positive contribution to the output and regions in blue color represent a negative contribution to the output. The white regions do not affect the output in any way.

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

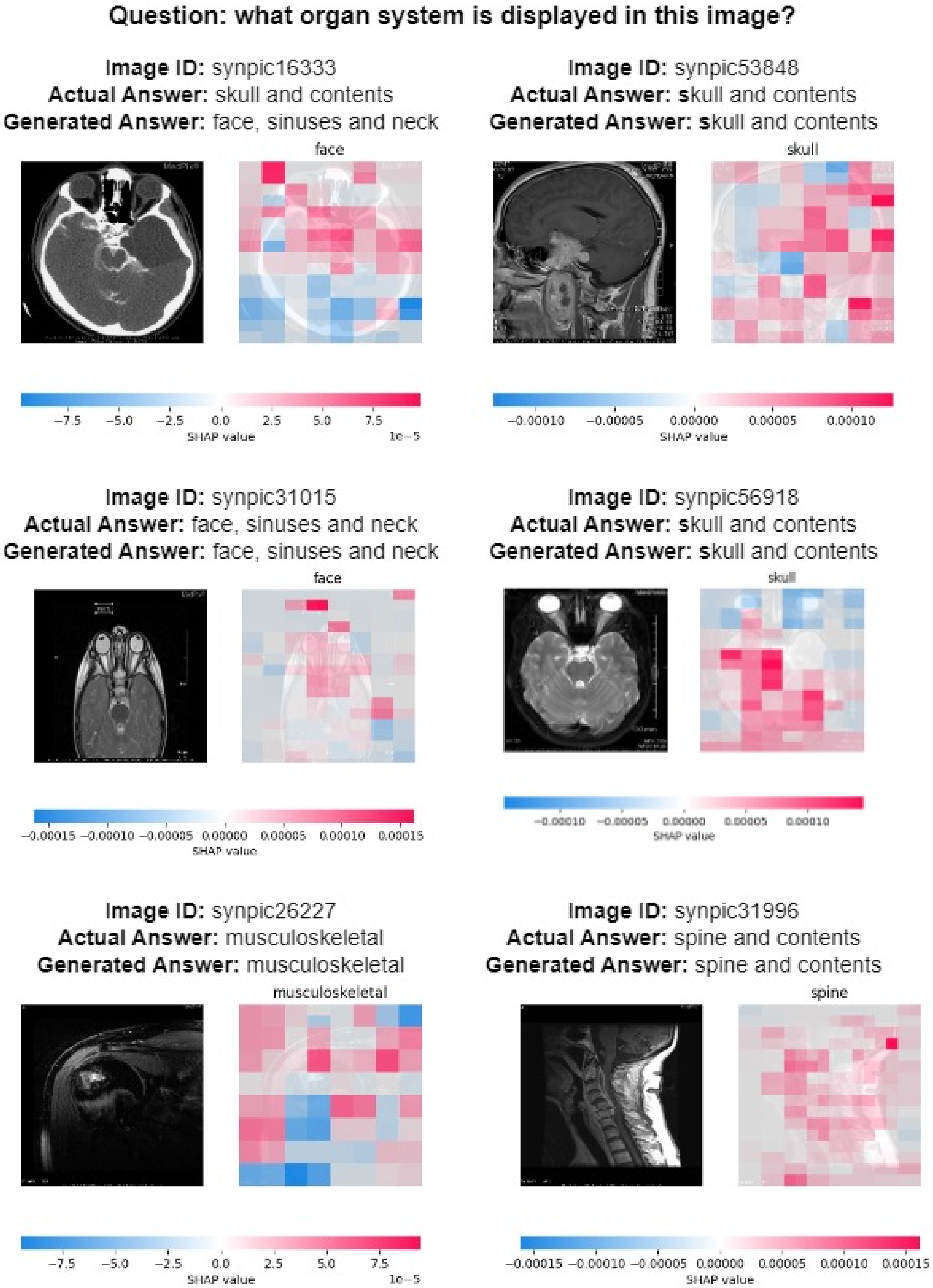


FIGURE 4.19: SHAP explanations for a set of sample images queried about organ

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.

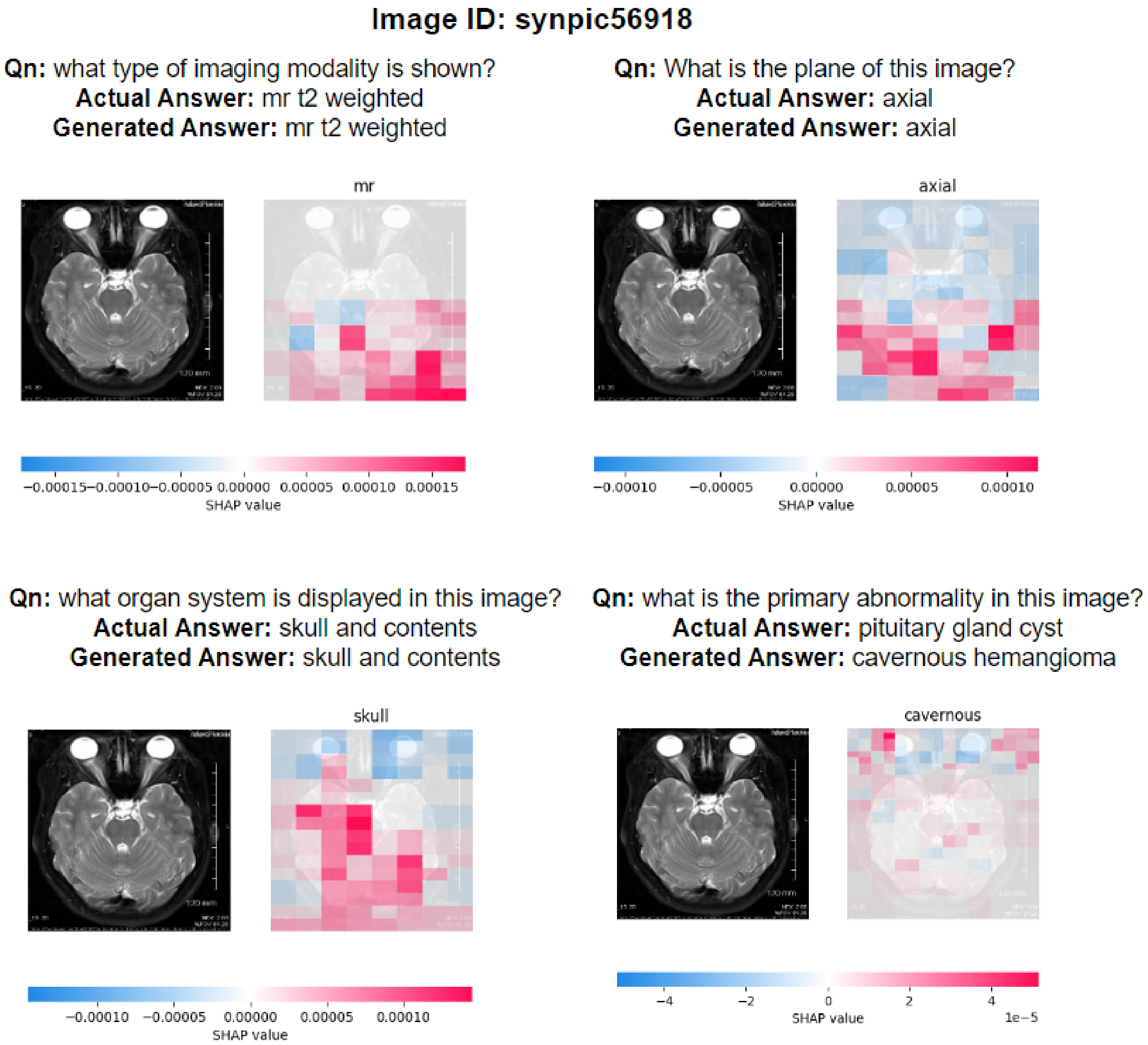


Image ID: synpic56918
Qn: what type of imaging modality is shown?
Actual Answer: mr t2 weighted
Generated Answer: mr t2 weighted
mr
−0.00015 −0.00010 −0.00005 0.00000 0.00005 0.00010 0.00015
SHAP value
Qn: What is the plane of this image?
Actual Answer: axial
Generated Answer: axial
axial
−0.00010 −0.00005 0.00000 0.00005 0.00010
SHAP value
Qn: what organ system is displayed in this image?
Actual Answer: skull and contents
Generated Answer: skull and contents
skull
−0.00010 −0.00005 0.00000 0.00005 0.00010
SHAP value
Qn: what is the primary abnormality in this image?
Actual Answer: pituitary gland cyst
Generated Answer: cavernous hemangioma
cavernous
−4 −2 0 2 4
SHAP value
1e−5

FIGURE 4.20: SHAP explanations for the model's working for different types of question

4.6 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 category 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

The ImageCLEF 2019 VQA-Med task winner's model resulted in 62.4% accuracy & 64.4 BLEU Score and the proposed VQA model resulted in 46.8% accuracy & 48.61 BLEU Score.

Discussion

The developed VQA model is validated using quantitative metrics and XAI techniques. Using quantitative metrics like accuracy, BLUE Score and WBSS are calculated for the ImageClef 2019 VQA-Med Dataset and compared with the task participants' results. The proposed VQA model resulted in less accuracy than the task winner since the combination of chosen techniques (VGGNet + BERT) is not suitable for abnormality-related questions which has more number of classes (1484 classes).

XAI techniques LIME and SHAP are applied to analyze the predicted outcome. Both LIME and SHAP modify the input images by inpainting some segments or partitions and computes the impact of the modifications on the output of the model. In the case of LIME, a local surrogate model is built for generating explanations. In SHAP, the impact of modifications on the input is measured in terms of Shapley values. Since LIME builds a local interpretable model using the perturbed dataset, the explanations generated are different for each turn depending upon the perturbed dataset created. Hence, the output of LIME is not consistent. Also, for some input instances, there are no explanations generated. This is also because of the perturbed dataset created by LIME. If there are not many samples in the perturbed dataset to identify the segments that affect the output of the model, LIME cannot generate explanations. LIME takes more time to generate explanations and it depends on the number of samples in the perturbed dataset.

SHAP also creates a set of samples by modifying the input. But it does not create a local model. It calculates the Shapley values for the set of features which represents the contribution of that set of features towards the output. Later the Shapley values of the set of features are averaged to get the global contribution of each feature. SHAP is comparatively faster than LIME and the explanations do not depend on the samples created by modifying the input. Hence the explanations are consistent and can give explanations for any input samples, unlike LIME.

CHAPTER 5

SOCIAL IMPACT AND SUSTAINABILITY

5.1 SOCIAL IMPACT

The medical domain requires faster and more efficient analysis of patients' conditions to diagnose and provide necessary treatments. Artificial Intelligence (AI) in healthcare would help to achieve such faster and efficient analysis using different kinds of data used in diagnosis. The data includes patients' symptoms, clinical test results and medical images. Medical images acquired from different modalities in general have low resolution and complex algorithms are required to analyze them. This project is proposed for developing a VQA system with XAI uses low-resolution images with different modalities from CLEF2019 VQA-Med dataset. The VQA system takes an input medical image and answers the questions related to the image. Further, the proposed project interprets the outcome i.e., the answer generated by the VQA model using Explainable AI (XAI) techniques.

These VQA systems that can give explanations can be integrated as a part of Medical Domain Expert Systems. These expert systems can be used not only by patients but the doctors can also use these systems when they interact with patients through online mode. Especially during pandemic situations, when patients prefer not to consult doctors in a physical mode, the VQA systems can be used for analyzing the medical images and diagnosing the medical conditions with interpretations.

5.2 SUSTAINABILITY

The proposed project uses the medical images from the ImageCLEF dataset to create a VQA model. The project can be extended to use other types of images like natural images, remote sensing images, etc., for building VQA systems. The XAI techniques LIME and SHAP that are used in this project for providing justifications are both model-agnostic techniques. Both LIME and SHAP can interpret the outcomes or predictions of any complex model.

The developed VQA + XAI system can work in any operating system like Windows, Ubuntu and Mac that has Python runtime with necessary libraries and software specified in Section 4.2.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This project is proposed for building a Visual Question Answering (VQA) model using the ImageCLEF 2019 VQA-Med dataset and analyzing the outcome of the model using Explainable AI (XAI) techniques. The dataset consists of 3200 images and 12792 questions for training, 500 images and 2000 questions for validating, and 500 images and 500 questions for testing the developed model. The image features are extracted using modified VGG-16, and the question features are extracted by tokenizing the question using the Bidirectional Encoder Representations from Transformers (BERT) tokenizer. A BERT model is trained to predict 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 the test dataset. Overall accuracy is less because abnormality-based questions has more number of classes (1484 classes). XAI techniques LIME and SHAP are used to generate explanations and justify the outcome of the VQA model. LIME and SHAP techniques generate explanations by finding segments of the image that contribute to predicted outcome. The explanations are visualized by inpainting the original input image.

In the future, an improved VQA model can be built and analysed for the ImageCLEF 2019 VQA-Med dataset using Global Average Pooling (GAP), attention modules, ResNet, DenseNet, LSTM, etc., for extracting features. Also, other VQA medical datasets with abnormality-related images, like PathVQA, ImageCLEF 2020 VQA-Med, etc., can be considered for model creation to improve accuracy. In addition, XAI techniques like Gradient-weighted Class Activation Mapping (Grad-CAM), Anchor, etc., can be explored by integrating with the VQA model for more 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, `https://ceur-ws.org/Vol-2380/paper_167.pdf`.

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, `https://ceur-ws.org/Vol-2380/paper_125.pdf`.

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, https://ceur-ws.org/Vol-2380/paper_272.pdf.

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, https://ceur-ws.org/Vol-2380/paper_117.pdf.

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. Gunning D and Aha D, (2019), DARPA's Explainable Artificial Intelligence (XAI) Program, AI Magazine, Vol. 40, Issue 2, pp. 44-58, https://doi.org/10.1609/aimag.v40i2.2850.

12. 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, https://ceur-ws.org/Vol-2380/paper_124.pdf.

13. 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.

14. 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.

15. 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, https://doi.org/10.18653/v1/N16-3020.

16. 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

17. 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, https://ceur-ws.org/Vol-2380/paper_64.pdf.

18. 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.

19. 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, https://doi.org/10.48550/arXiv.1705.07874.

20. 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, https://ceur-ws.org/Vol-2380/paper_127.pdf.

21. 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, https://ceur-ws.org/Vol-2380/paper_123.pdf.

22. 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.

23. 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.

24. 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, https://ceur-ws.org/Vol-2380/paper_85.pdf.

25. 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, https://ceur-ws.org/Vol-2380/paper_190.pdf.

26. 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.

27. 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, https://doi.org/10.48550/arXiv.2111.10056.

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

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