

**VISUAL QUESTION ANSWERING FOR RADIOLOGY IMAGE
OF SKELETAL SCINTIGRAPHY IN MEDICAL DOMAIN USING
FEATURE EXTRACTION METHOD**

रेडियोलॉजी छवि के लिए दृश्य प्रश्न उत्तर चिकित्सा क्षेत्र में स्केलेटल स्कॅन्टिग्राफी
का उपयोग फीचर निष्कर्षण विधि

A

Thesis

**Submitted for the Award of the Ph.D. degree of
PACIFIC ACADEMY OF HIGHER
EDUCATION AND RESEARCH UNIVERSITY**

By

JINESH MELVIN Y I

जिनेश मेल्विन वाई आई

Under the supervision of

Dr. MUKESH SHRIMALI

Professor,
Pacific Academy of Higher
Education & Research University, Udaipur

Dr. SUSHOPTI GAWADE

Professor,
PCE, Department of Computer Engineering,
New Panvel, Navi Mumbai



**FACULTY OF ENGINEERING
DEPARTMENT OF COMPUTER ENGINEERING
PACIFIC ACADEMY OF HIGHER EDUCATION
AND RESEARCH UNIVERSITY, UDAIPUR
2024**

DECLARATION

I, **JINESH MELVIN Y I S/O SHRI YOHANNAN** resident of Sankalp Shiddi, Plot No. 59, Sector 5A, Room 1003, Karanjade, Panvel, 410206, hereby declare that the research work incorporated in the present thesis entitled “**Visual Question Answering for Radiology Image of Skeletal Scintigraphy in Medical Domain Using Feature Extraction Method**” (रेडियोलॉजी छवि के लिए दृश्य प्रश्न उत्तर चिकित्सा क्षेत्र में स्केलेटल स्किंटिग्राफी का उपयोग फीचर निष्कर्षण विधि) is my original work. This work (in part or in full) has not been submitted to any University for the award or a Degree or a Diploma. I have properly acknowledged the material collected from secondary sources wherever required.

I solely own the responsibility for the originality of the entire content.

Signature of the Candidate

Date:

FACULTY OF ENGINEERING

PACIFIC ACADEMY OF HIGHER EDUCATION AND RESEARCH UNIVERSITY, UDAIPUR

Dr. MUKESH SHRIMALI

Professor

CERTIFICATE

It gives me immense pleasure in certifying that the thesis “**Visual Question Answering for Radiology Image of Skeletal Scintigraphy in Medical Domain Using Feature Extraction Method**” (रेडियोलॉजी छवि के लिए दृश्य प्रश्न उत्तर चिकित्सा क्षेत्र में स्केलेटल स्कॅन्टिग्राफी का उपयोग फीचर निष्कर्षण विधि) and submitted by **JINESH MELVIN Y I** is based on the research work carried out under my guidance. He / she have completed the following requirements as per Ph.D. regulations of the University;

- (i) Course work as per the University rules.
- (ii) Residential requirements of the University.
- (iii) Regularly presented Half Yearly Progress Report as prescribed by the University.
- (iv) Published / accepted minimum of two research paper in a refereed research journal.

I recommend the submission of thesis as prescribed/notified by the University.

Date:

Name and Designation of Supervisor

Dr. MUKESH SHRIMALI

Professor,
Pacific Academy of Higher Education
& Research University, Udaipur

CERTIFICATE

It gives me immense pleasure in certifying that the thesis “**Visual Question Answering for Radiology Image of Skeletal Scintigraphy in Medical Domain Using Feature Extraction Method**” (रेडियोलॉजी छवि के लिए दृश्य प्रश्न उत्तर चिकित्सा क्षेत्र में स्केलेटल स्कॅन्टिग्राफी का उपयोग फीचर निष्कर्षण विधि) and submitted by **JINESH MELVIN Y I** is based on the research work carried out under my guidance. He / she have completed the following requirements as per Ph.D. regulations of the University;

- (i) Course work as per the University rules.
- (ii) Residential requirements of the University.
- (iii) Regularly presented Half Yearly Progress Report as prescribed by the University.
- (iv) Published / accepted minimum of two research paper in a refereed research journal.

I recommend the submission of thesis as prescribed/notified by the University.

Date:

Name and Designation of Co-Supervisor

Dr. SUSHOPTI GAWADE

Professor,
PCE, Department of Computer Engineering,
New Panvel, Navi Mumbai

COPYRIGHT

I, **JINESH MELVIN Y I**, hereby declare that the Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan, shall have the rights to preserve, use and disseminate this dissertation entitled **“Visual Question Answering for Radiology Image of Skeletal Scintigraphy in Medical Domain Using Feature Extraction Method”** (रेडियोलॉजी छवि के लिए दृश्य प्रश्न उत्तर चिकित्सा क्षेत्र में स्केलेटल स्कैन्टिग्राफी का उपयोग फीचर निष्कर्षण विधि) in print or in electronic format for the academic research.

Date:

Signature of Candidate

Place:

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my supervisor, **Dr. Mukesh Shrimali** and **Dr. Sushopti Gawade**, for their unwavering support, invaluable guidance, and immense patience throughout the journey of completing this doctoral thesis. Their expertise, encouragement, and constructive feedback have been instrumental in shaping the direction of my research and fostering my academic growth.

I am also grateful to the members of my doctoral committee members, for their valuable insights, thoughtful suggestions, and rigorous examination of my work. Their expertise and scholarly contributions have enriched the quality of this thesis.

I extend my heartfelt appreciation to **Dr. Hemant Kothari** from Pacific University, whose collaboration and assistance have contributed significantly to the success of this research project. Their expertise, collaboration, and willingness to share resources have been invaluable.

I am deeply indebted to my lovable parents specially, my Mom Ida Christabel, my Dad **Yohannan M**, my Wife **Ginchil Krus** and their parents **John Krus** and **Lyla Christal** for their unwavering love, encouragement, and support throughout this challenging yet rewarding journey. Their patience, understanding, and belief in me have been a constant source of strength and motivation.

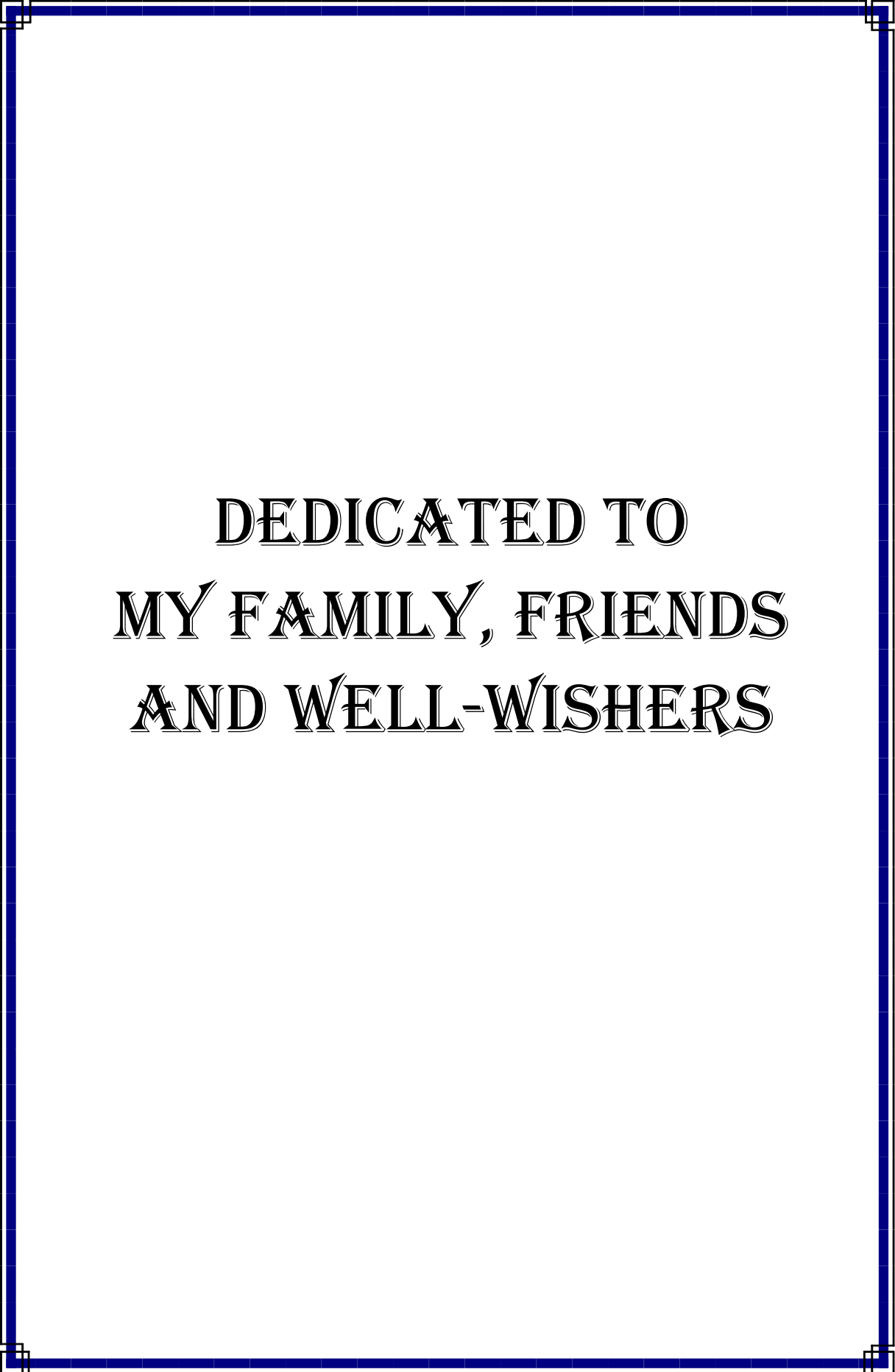
I would like to acknowledge the support and resources provided by Pillai College of Engineering and Pacific University, without which this research would not have been possible. The facilities, resources, and academic environment provided by the university have been crucial to the completion of this thesis.

Finally, I wish to express my gratitude to all the individuals who have contributed in any way, no matter how small, to the completion of this thesis. Your support, encouragement, and inspiration have been deeply appreciated

The final one, my distinctive thanks to *Nav Nimantran Thesis Printing & Binding, Udaipur* Admin Team for their role in shaping this research, creative design work and bringing out this document meticulously, neatly and timely.

DATE: -

JINESH MELVIN Y I



DEDICATED TO
MY FAMILY, FRIENDS
AND WELL-WISHERS

INDEX

CHAPTER- I INTRODUCTION		1 – 19
1.1	Common discussion about VQA in healthcare domain	1
1.2	Image and Textual feature representations	2
1.3	Question Answer Processing	3
1.4	Different Question Types with corresponding Answers	5
1.5	Motivation of VQA in healthcare domain	6
1.6	Scope of VQA system in the healthcare domain	7
1.7	Hypothesis of VQA system in healthcare industry	9
1.8	Research contribution to the society	12
1.9	Application of VQA	14
1.10	Organizations of Thesis	15
1.11	Problem Statement for Visual Question Answering System	16
1.12	Objectives for VQA's proposed approach	17
CHAPTER- II REVIEW OF LITERATURE		20 - 52
2.1	Review of relevant literature and previous research	20
	2.1.1 Review on Visual Question Answer	20
	2.1.2 Review on Radiology Image Datasets	25
	2.1.3 Research on Current Methodology with existing techniques and Algorithms	31
	2.1.4 Survey on visual and textual Feature Extraction Technique	37
2.2	Identification of challenges and gaps in existing research	46
	2.2.1 Challenges in VQA System	46
	2.2.2 Inhibitions of Datasets	47
	2.2.3 Drawbacks on various techniques	48
2.3	Gap Identification from Existing Research	49
CHAPTER-III METHODOLOGY		53 – 102
3.1	Description of the research design and methodology used	53
	3.1.1 Morphology of a Radiology Image	55
3.2	Explanation of data collection methods, tools, and procedures	57

3.3	Current Methodology on Both Visual and Textual Feature Extraction Techniques	58
	3.3.1 Linear Classifier	58
	3.3.2 Traditional Neural Network	59
	3.3.3 The prime Idea behind a Convolutional Neural Network for image feature extraction	60
	3.3.4 Deep Belief Network model, the feature extraction of skeletal image	74
	3.3.5 The deep insights of Region-based Convolutional Neural Network (R-CNN) for visual classification	80
	3.3.6 Faster Region-based Convolutional Neural Network (R-CNN)	84
	3.3.7 Long Short-Term Memory networks for question answering system	90
3.4	The proposed Approach for Visual Question Answering System (VQAS) for skeletal Images	93
	3.4.1 Skeletal image Feature Extraction using Block_12_add Faster R-CNN (B12- FRCNN) algorithm	94
	3.4.2 The proposed approach Kai-Bi-LSTM for Question Answering feature extraction	98
CHAPTER- IV EXPERIMENTAL RESULT AND ANALYSIS		103 - 139
4.1	Experimental Result on VQA	103
4.2	Code Implementation	117
4.3	Comparison analysis of Feature extraction techniques	120
4.4	Snapshot on demonstration	125
CHAPTER- V CONCLUSION AND FUTURE ENHANCEMENT		140 – 142
REFERENCE		143 - 155
PUBLICATIONS		
CERTIFICATES		
PLAGIARISM REPORT		

LIST OF TABLE

Table No.	Particulars	Page No.
1	Descriptive Statistics of Existing visual and textual dataset	28
2	Correlation of Existing visual and textual dataset	28
3	Covariance of Existing visual and textual dataset	29
4	Cumulative Frequency of Existing visual and textual dataset	29
5	Describes the publications obtained and reviewed during the current survey	41
6	Most Frequent Answers for various Question Type and Answer count	110
7	Total count of Visual and Textual dataset	112
8	The table presents the performance metrics for a certain method across different evaluation criteria	120
9	The accuracy achieved by different classification algorithms	122
10	The table shows the F-measure achieved by different algorithms	123
11	The table compares the accuracy of both Proposed and Current algorithms	124
12	The comparative analysis between Existing CNN and Existing BiLSTM with Proposed BiLSTM	125

LIST OF FIGURE

Fig. No.	Particulars	Page No.
1	Visual Question Answering system with medical image	1
2	Trends in Visual Question Answering Research	21
3	Histogram of existing collective dataset with its frequency	30
4	Total data present in various image and question answer pair dataset	30
5	Insight of Literature Survey	45
6	Types of Radiology Image	56
7	Traditional Neural Network with a Single Layer	60
8	RGB Image and Gray scale Image to find features of the image	60
9	3×3 pixel object	61
10	Structure of CNN process from Input object to final output object with feature extraction	61
11	The value of Black and white colors 0 to 255	61
12	How the convolution operation take place with 6×6 Pixel images and 3×3 image	62
13	Black and white input image for convolution operation to identify the vertical edge	62
14	Black and White image with 3×3 px vertical edge filter the output is in 4×4 px	63
15	Output image from above figure after 3×3 px convolution with 6×6 px	63
16	Image 3×3 px horizontal edge filter	64
17	$n \times n \times 1$ for Gray scale image with $f \times f \times c$ number of filters the output will be multiple images	64
18	Single filter convolution with single output	66
19	Multiple filter convolution with multiple output	66

Fig. No.	Particulars	Page No.
20	Padding in Convolutional Neural Network	66
21	Single Filter Convolution layer using non linear function and bias with single image as output	69
22	Multiple Filter Convolution layer using non linear function and bias with multiple output So, in short, we convert it as	70
23	Overall Convolution layer using non linear function and bias with multiple output in one frame	71
24	The Architecture of Convolutional Neural Network	73
25	The Traditional Architecture of Artificial Neural Network	74
26	Two-layer probabilistic neural network, RBM Architecture	76
27	Overall architecture of a Deep Belief Network	78
28	Faster Region-based Convolutional Neural Network (R-CNN)	84
29	Understanding the Fast RCNN and Faster RCNN Architecture	90
30	Architecture of Long Short-Term Memory (LSTM) networks	93
31	B12-FRCNN, the "block_12_add" layer	97
32	Block diagram of QA System in the Training phase	102
33	Block diagram of QA System on Testing Phase	102
34	Confusion Matrix	106
35	Most frequent answers from different question types	111
36	Different methods of question for various question types	112
37	Total count of Image and questions from CLEF Image Retrieval and Classification Task 2019 Dataset	112
38	Total number of data from both question and answer for various types	114
39	Count of boolean questions for Modality	114
40	Various types of question for Modality question answering data type	115

Fig. No.	Particulars	Page No.
41	Various types of question for Plane question answering data type	115
42	Various types of question for Organ question answering data type	116
43	Various types of question for Abnormality question answering data type	116
44	Overall count of different question types	116
45	Sample code function on Existing CNN algorithm	117
46	Construction of Existing DBN structure for Medical Images	118
47	Code function on Proposed Kai_BiLSTM	120
48	The performance metrics for a certain method across different evaluation criteria	121
49	The accuracy achieved by various classification techniques for current methodology	122
50	F-measure achieved by different algorithms	123
51	Comparison with existing algorithm and proposed algorithm	124
52	The Comparison Result of Proposed and Existing algorithms	125
53	Starting page of Application	125
54	Load the Training dataset both Image and Question Answer pair	126
55	Feature Extraction for Image dataset	126
56	Pre-process the Question dataset	127
57	Pre-process for Answer Dataset	127
58	Done with Preprocessing for both Question and Answer dataset	128
59	Word Embedding for Question Dataset using BERT model	128

Fig. No.	Particulars	Page No.
60	Word Embedding for Answer Dataset using LBER model	129
61	Training the dataset for both visual and textual dataset	129
62	Load the Testing dataset both Image and Question Answer pair	130
63	Feature Extraction for Testing image dataset	130
64	Pre-processing for Question Datasets using LBER model	131
65	Testing both visual and textual dataset	131
66	Precision measure for proposed and existing models	132
67	Recall measure for proposed and existing models	132
68	Classification of FMeasure for proposed and existing models	133
69	Classification Sensitivity measure for proposed and existing models	133
70	Classification Specificity measure for proposed and existing models	134
71	True Positive Rate (TPR) measure for proposed and existing models	134
72	Positive Predictive Value (PPV) measure for proposed and existing models	135
73	Measure False Negative Rate (FNR) for proposed and existing models	135
74	To display Graphs and Tables	136
75	Load the skeletal Image	136
76	Feature Extraction for Loaded Image using B12-FRCNN	137
77	Users Input Question	137
78	Preprocess the Input Question	138
79	Word Embedding for Input Question	138
80	Classification using Kai-BiLSTM model	139
81	Predicated answer with calculated score value for the answer	139

ABBREVIATIONS

ABBREVIATIONS	DEFINITIONS
AI	Artificial Intelligence
B12 FRCNN	Block 12 Faster Region-based Convolutional Neural Networks
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
BPI-MVQA	Benchmark for Performance Evaluation of Medical Visual Question Answering
BPTT	Backpropagation Through Time
BS	Base Station
CAT	Computerised Axial Tomography
CGMVQA	Classification and Generative Model for Medical Visual Question Answering
CHAOS	challenge aims the segmentation of abdominal organs
CLEVR	Compositional Language and Elementary Visual Reasoning
CNN	Convolutional Neural Network
Co-DUA	Coupled-Deep Unsupervised Autoencoder
CQA	Community Question Answering
CS	Cosine Similarity
CTS	Computed Tomography Scan
DBN	Deep Belief Network
De-DUA	Differential Evolutionary Deep Unsupervised Autoencoder
DL-UL	Downlink-Uplink
ED	Euclidean Distance
EHRs	Electronic Health Records
FC	Fully Connected layer
FN	False Negative
FP	False Positive
GELU	Gaussian Error Linear Unit
GI	gastrointestinal
GloVe	Global Vectors for Word Representation
GPT	Generative Pre-trained Transformer

GRU	Gate Recurrent Unit
ImageCLEF	Cross Language Evaluation Forum
JSC	Jaccard Similarity Coefficient
Kai-BiLSTM	Kaimming Bidirectional long short-term memory
LSTM	long Short-Term memory
MD	Manhattan Distance
MFB	multi-modal factorized bilinear pooling model
MFN	Medical Fusion Network
Microsoft COCO	Microsoft Common Objects in Context dataset
MIMIC_CXR	Medical Information Mart for Intensive Care - Chest X-Ray
MLM	Masked Language Modeling
MLP	multi-layer perceptron
MQR-VQA	Multi-Modal Quality-Aware Relevance
MRA	Magnetic Resonance Angiography
MRI	Magnetic resonance imaging
MVQA	Medical Visual Question Answer
NIH	National Institutes of Health
NLP	Natural Language Processing
NMS	Non-maximum suppression
NSP	Next Sentence Prediction
PA	posterior-anterior
PCD	Persistent Contrastive Divergence
PEIR	Pattern Extraction and Identification Resource
PET	positron emission tomography
QA	Question Answer
QG	Question Generation
R-CNN	Region-based Convolutional Neural Network
RadVisDial	Radiology Visual Dialog dataset
RBM	Restricted Boltzmann Machine
ReLU	Rectified Linear Unit
ResNet	Residual Network
RFA	reverse frequency allocation

RGB	Red Green, and Blue
RNN	Recurrent Neural Network
RoI	Region of Interest
RPNs	region proposal networks
SLAKE	Semantically-Labeled Knowledge-Enhanced dataset
SSD	Single Shot MultiBox Detector
SSM	Skeleton-based Sentence Mapping
ST-VQA	Spatial-Temporal Visual Question Answering
SVM	Support Vector Machine
TP	True Positive
VGG16	Visual Geometry Group 16
ViT	Vision Transformer
VQA	Visual Question Answer
VQA-Med	Visual Question Answer Medical dataset
VQA-RAD	Visual Question Answering in Radiology
VQG	Visual Question Generation
Word2Vec	Word to Vector
YOLO	You Only Look Once

PREFACE

The Visual Question Answering (VQA) system represents an innovative fusion of computer vision and natural language processing, facilitating the capability of machines to respond to queries grounded in visual stimuli. In this scholarly task, we present a tailored VQA framework meticulously crafted for skeletal imagery, leveraging sophisticated techniques for extracting both visual and textual features. These techniques enable the system to comprehend the structural aspects of the skeletal system and gain insights from radiographic or medical imaging data. By effectively transforming the questions into textual features, the system gains a deeper understanding of the user's inquiries and can provide accurate answers. The fusion of both visual and textual features is achieved using sophisticated integration methods, ensuring a seamless correlation between the image content and the textual context. This integration empowers the system to reason effectively and formulate responses that are contextually relevant and adequate. To examine the effectiveness of our proposed VQA system, we conducted extensive experiments on a diverse dataset of skeletal images and corresponding textual queries. The results demonstrate the system's capability to provide accurate and insightful answers, showcasing its potential for applications in the healthcare domain, radiology of skeletal images, and beyond.

A novel skeletal image of the proposed approach is based on B12 FRCNN and Kai-Bi-LSTM approaches is introduced in this paper to address the different challenges. The proposed system aims to enhance communication between medical professionals and patients by providing accurate answers to visual questions related to medical images. The system uses advanced methods like B12 FRCNN for object localization and Kai-Bi-LSTM for sequential processing to try to understand and interpret medical image queries better. This should lead to better interactions between patients and doctors.

INTRODUCTION



INTRODUCTION

The proposed approach represents a dynamic and interdisciplinary domain situated at the convergence of computer vision and natural language processing (NLP). It represents a significant advancement in artificial intelligence, enabling machines to comprehend visual information and answer questions related to images, videos, or any visual content.

The idea behind VQA is to give intelligent machines the ability to interpret and reply to questions in natural language regarding the contents of the image in question. This integration of vision and language opens up a plethora of practical applications, ranging from robotics and autonomous systems to assistive technology development and vision impairment accessibility.

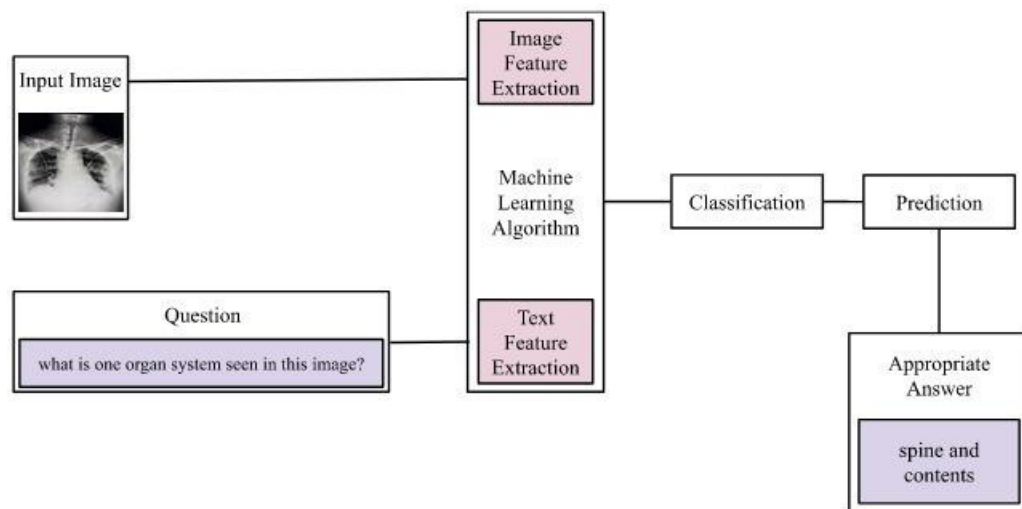


Fig. 1: Visual Question Answering system with medical image

1.1 Common discussion about VQA in healthcare domain

The primary aspiration is to develop a robust system adept at comprehending and addressing inquiries pertaining to bone characteristics and irregularities. Leveraging leading-edge for skeletal image representation algorithms, we emphasize on extracting visual features to encode the distinctive traits and patterns observed in skeletal images. Simultaneously, on the textual front, we employ advanced natural language processing algorithms to effectively process and extract meaningful insights from textual queries inputted into the Visual Question Answering (VQA) system. Recognizing that optimal communication between healthcare professionals and patients is paramount in obtaining accurate information concerning bodily conditions

and overall health, our endeavors aim to facilitate seamless interaction and understanding within this domain. Engaging in discussions concerning health matters may pose challenges for patients when faced with the technical terminology commonly employed by healthcare professionals. Successfully addressing patient concerns thus requires a multifaceted skill set, encompassing object localization, attribute identification, scene comprehension, reasoning, and counting. Attaining the requisite level of accuracy in these endeavors typically hinges on the availability of extensive labeled datasets, upon which most supervised learning algorithms heavily rely.

The fundamental challenge in VQA lies in uniting the space between the different modalities of information: visual data represented by pixels and textual data represented by natural language. To address this challenge, VQA systems employ sophisticated techniques for both visual and textual feature extraction.

1.2 Image and Textual feature representations

Image Feature Extraction: At the heart of VQA is the process of extracting significant information from healthcare skeletal images or videos. Convolutional Neural Networks (CNNs) are commonly used for visual feature extraction, allowing the system to capture intricate patterns, objects, and visual context. These visual features are then encoded into a compact representation that captures the essence of the image's content.

Textual Feature Extraction: To comprehend and process the questions, VQA systems employ natural language processing techniques. Recurrent neural networks (RNNs) or transformer-based models are frequently used to convert textual input into meaningful embeddings. These textual features encode the semantics and context of the question, enabling the system to understand the user's inquiry.

Integration of Visual and Textual Information: Once the Image and textual features are extracted, the VQA system fuses these representations to build a cohesive understanding of the input. Various fusion mechanisms, such as attention mechanisms or multimodal embeddings, are used to align the visual and textual domains effectively. This fusion facilitates the correlation between the question and the visual content, allowing the system to reason and generate appropriate answers.

Applications of Visual Question Answering: VQA systems find application in numerous real-world scenarios. For instance, they can assist autonomous vehicles in understanding their surroundings and responding to spoken instructions. In healthcare, VQA systems can aid medical professionals in diagnosing medical images or assist visually impaired individuals in understanding visual content.

Challenges and Future Directions: While significant progress has been made in VQA, challenges persist, such as handling complex questions, handling ambiguous queries, and ensuring robustness to noisy input. Researchers continue to explore novel approaches, leveraging pre-training, multimodal reasoning, and larger datasets to improve VQA system performance.

1.3 Question Answer Processing

This system incorporates interrogative response systems to assist patients, as depicted in Fig. 1. To achieve accurate output prediction, a substantial amount of data needs to be trained, including Enormous datasets containing subjective and descriptive answers to numerous questions.

However, dealing with such massive datasets poses challenges due to the computational resources required to process them effectively. Thus, a critical aspect of the system is to reduce the size of data without losing any important information. This data reduction process is crucial for optimizing machine efforts, speeding up performance, and enhancing the system's learning capabilities.

The system employs feature extraction techniques to store high-quality datasets in a new location. Utilizing pattern analysis techniques, particularly the classification algorithm, the system can classify and To accurately forecast the output answer based on the question posed about the image is to efficiently interpret and derive the most suitable response aligned with the inquiry.

During testing, the system takes an image and processes various questions to predict the appropriate answer using the data mining classification algorithm. If necessary, the system may suggest related images to further clarify the input images and questions, ensuring better support and understanding for patients using the VQA system.

Thousands of images are uploaded for training in the Visual Question Answering (VQA) system, along with corresponding relevant questions and answers for each individual image. It is common to have multiple questions and answers associated with a single image, capturing different aspects and interpretations of the visual content.

During the training process, the VQA system learns to map the images to their respective question-answer pairs, enabling it to comprehend the association between the visual features and textual information. By analyzing diverse question-answer combinations, the model acquires a more comprehensive knowledge of the image context and improves its ability to generate appropriate responses.

Having multiple questions and answers for a single image enhances the system's capability to handle variations in phrasing, question structure, and answer styles. This variability in the training data helps the VQA system generalize better to real-world scenarios where users may express their queries differently.

To achieve optimal performance, it is essential to curate a diverse and well-balanced dataset, ensuring a wide representation of questions and answers that cover various aspects of the images. This data diversity elevates the overall functionality and versatility of the VQA system in addressing a broad spectrum of visual inquiries.

In a specialized Visual Question Answering (VQA) system designed for medical images, users can pose diverse questions about the content of the images. These questions encompass a range of medical inquiries. For instance, users can ask about the identification of specific anatomical structures or abnormalities, measurements of sizes, potential diagnoses based on visual findings, recommended treatments, imaging modalities used, disease severity assessments, common medications, and treatment response evaluations.

The system aims to generate accurate and contextually relevant answers to aid healthcare professionals in medical diagnosis, treatment planning, and patient care. To achieve this, the VQA system relies on a comprehensive knowledge base of medical data, allowing it to respond with precision and provide valuable insights into the medical images. Careful validation of the answers by medical experts ensures the

appropriateness and reliability of the information provided, making the VQA system a valuable tool in the medical domain.

1.4 Different Question Types with corresponding Answers

In a Visual Question Answering (VQA) system tailored for medical images, users can ask various types of questions about the content of the images. Here are some common question types and their corresponding types of answers in a medical VQA system:

1. Anatomical identification questions:
 - Question: "What part of the body is shown in the image?"
 - Answer: A label representing the anatomical structure or body part depicted in the medical image (e.g., "liver," "lung").
2. Abnormality Detection Questions:
 - Question: "Is there any abnormality or pathology in the image?"
 - Answer: A binary response indicating the positive or negative of any medical abnormality (e.g., "Yes" or "No").
3. Quantitative Measurement Questions:
 - Question: "What is the size of the tumor in the image?"
 - Answer: A numerical value representing the size or measurement of a specific anatomical structure or pathology (e.g., "5 cm").
4. Differential Diagnosis Questions:
 - Question: "What could be the possible diagnosis based on the image?"
 - Answer: A list of potential medical conditions or diagnoses that could be associated with the visual findings in the image.
5. Treatment or Intervention Questions:
 - Question: "What is the recommended treatment for this condition?"
 - Answer: A description or a list of treatment options or medical interventions for the identified medical condition.
6. Image Modality Questions:
 - Question: "Is this image obtained from an X-ray or an MRI?"
 - Answer: A label representing the imaging modality used to acquire the medical image (e.g., "MRI").
7. Disease Severity Questions:

- Question: "How severe is the disease shown in the image?"
 - Answer: A qualitative descriptor or a numerical score indicating the severity of the identified medical condition.
8. Medication Questions:
- Question: "What medications are typically prescribed for this condition?"
 - Answer: A list of common medications used in the treatment of the identified medical condition.
9. Treatment Response Questions:
- Question: "Is the patient's condition improving after treatment?"
 - Answer: A qualitative response indicating the response of the patient's condition to a specific treatment.

It's important to note that medical VQA systems require specialized knowledge and access to accurate and reliable medical databases to provide accurate and relevant answers. The answers generated by the system should be well-vetted by medical professionals to ensure their accuracy and appropriateness for clinical decision-making. Such systems have competence to provide support to healthcare professionals in medical diagnosis, treatment planning, and patient care.

1.5 Motivation of VQA in healthcare domain

To construct a Visual QA (VQA) system in the healthcare industry is motivated by several key factors:

1. **Improved Patient Care:** A VQA system improves the healthcare outcomes and personnel make better decisions about patient care by offering easy access to essential information. For example, doctors can ask questions about medical images, such as X-rays or MRIs, to obtain insights about a patient's condition, potential diagnoses, or treatment options.
2. **Efficient Diagnosis:** Healthcare professionals often need to examine the large volumes of skeletal images to diagnose diseases or identify abnormalities. A VQA system can help streamline this process by automatically extracting relevant information from images and answering queries posed by clinicians. This can result in faster and more accurate diagnoses, ultimately improving patient outcomes.

3. **Enhanced Medical Education:** VQA systems have the potential to serve as effective educational aids for medical students, residents, and other healthcare professionals. By allowing users to ask questions about the skeletal images and receive informative responses, these systems can facilitate learning and knowledge retention in a more interactive and engaging manner.
4. **Remote Consultations:** In scenarios where specialists may not be physically present, such as in rural or underserved areas, a VQA system can enable remote consultations between healthcare providers. Clinicians can share medical images and ask questions to obtain expert opinions and guidance, leading to better patient management and care coordination.
5. **Research and Innovation:** VQA systems can support medical research efforts by providing access to huge datasets of annotated skeletal images. Researchers can use these systems to pose research questions, analyze image data, and gain insights into various medical conditions and treatments. Additionally, VQA systems can foster innovation by enabling the development of advanced image analysis algorithms and machine learning models.

Overall, the development of VQA systems in the healthcare industry is driven by the goal of improving patient care, enhancing medical education, facilitating remote consultations, and advancing medical research and innovation. By leveraging the power of computer vision and natural language processing technologies, these systems have the potential to transform various aspects of healthcare delivery and contribute to better health outcomes for patients worldwide.

1.6 Scope of VQA system in the healthcare domain

The scope of developing a Visual QA (VQA) system in the healthcare domain is vast and encompasses numerous opportunities for innovation and improvement in patient care, medical education, research, and more. Below are some key aspects that illustrate the scope of VQA systems in healthcare:

1. Diagnostic Support:

- VQA systems can assist healthcare professionals in diagnosing medical conditions by providing additional context and insights from medical images.

- Clinicians can ask questions about specific features or anomalies in images, and the VQA system can generate responses based on its understanding of the images.

2. Treatment Planning and Decision Support:

- VQA systems can aid in treatment planning by providing relevant information about treatment options, potential side effects, and patient outcomes based on similar cases.
- Healthcare providers can ask questions about treatment protocols, drug interactions, and surgical procedures, and the VQA system can provide evidence-based recommendations.

3. Medical Education and Training:

- VQA systems can serve as valuable educational assets for healthcare professionals in training, residents, and other healthcare professionals by providing interactive learning experiences.
- Learners can ask questions about skeletal images and receive detailed explanations, helping them understand complex medical concepts and procedures.

4. Remote Consultations and Telemedicine:

- VQA systems can facilitate remote consultations between healthcare providers and patients, remarkably in underserved or remote areas where access to specialists may be constrained.
- Patients can share medical images with their healthcare providers and ask questions about their condition, treatment options, and follow-up care.

5. Research and Data Analysis:

- VQA systems can support medical research efforts by providing access to huge datasets of annotated skeletal images and associated clinical data.
- Researchers can use VQA systems to analyze image data, extract meaningful insights, and identify patterns or correlations that may inform new research directions or treatment strategies.

6. Patient Engagement and Empowerment:

- VQA systems can empower patients by providing them with a better understanding of their medical conditions and treatment plans.
- Patients can ask questions about their medical images, laboratory results, or treatment options, and the VQA system can provide clear and understandable explanations, helping patients make informed decisions about their health.

7. Quality Improvement and Clinical Decision Support:

- VQA systems can contribute to quality improvement initiatives by assisting healthcare providers in making more accurate and timely clinical decisions.
- By analyzing large volumes of medical image data and providing relevant information in real time, VQA systems can help clinicians optimize their workflow and improve patient outcomes.

The scope of developing VQA systems in healthcare is broad and multifaceted, encompassing various aspects of patient care, medical education, research, and quality improvement. By leveraging advances in computer vision, natural language processing, and machine learning, VQA systems have the potential to revolutionize healthcare delivery and contribute to better health outcomes for patients worldwide.

1.7 Hypothesis of VQA system in healthcare industry

The hypothesis of a Visual QA (VQA) system in the healthcare industry posits that integrating computer vision and natural language processing techniques can significantly enhance medical image analysis, clinical decision-making, patient engagement, and healthcare outcomes.

- 1. Improved Diagnostic Accuracy:** The VQA system can assist healthcare providers in accurately interpreting medical images by answering specific questions related to visual patterns, abnormalities, and diagnostic criteria. By leveraging both visual and textual information, the system aims to enhance diagnostic accuracy and reduce errors in image interpretation.
- 2. Enhanced Clinical Decision Support:** Through intelligent analysis of skeletal images and contextual understanding of clinical questions, the VQA system can provide timely and personalized decision support to healthcare

professionals. This support includes treatment recommendations, differential diagnoses, and prognostic insights based on image features and patient data.

3. **Efficient Workflow Integration:** By seamlessly integrating into existing clinical workflows and electronic health record systems, the VQA system aims to streamline image interpretation, consultation, and reporting processes. This integration facilitates efficient communication between healthcare providers, reduces turnaround time, and improves overall workflow efficiency.
4. **Empowered Patient Engagement:** The VQA system empowers patients to take an active role in their healthcare journey by providing understandable explanations and insights into their medical images. Patients can ask questions, seek clarification, and make informed decisions about their treatment plans, leading to increased engagement, satisfaction, and adherence to therapy.
5. **Facilitated Medical Education and Training:** Medical students, residents, and healthcare professionals can leverage the VQA system as a valuable educational tool for learning about medical imaging interpretation, anatomy, pathology, and clinical reasoning. The system provides interactive learning experiences, case-based tutorials, and real-time feedback to support continuous professional development.
6. **Accelerated Research and Innovation:** Researchers and scientists can leverage the VQA system to analyze large-scale medical image datasets, identify novel biomarkers, and discover patterns indicative of disease progression, treatment response, and therapeutic efficacy. This accelerated research process enables the development of innovative diagnostic tools, predictive models, and precision medicine approaches.
7. **Enhanced Quality of Care and Patient Outcomes:** Ultimately, the hypothesis suggests that the adoption of VQA systems in healthcare can lead to improvements in diagnostic accuracy, clinical decision-making, patient engagement, workflow efficiency, medical education, research productivity, and, most importantly, patient outcomes. By harnessing the power of artificial intelligence and human expertise, VQA systems have the ability to completely transform healthcare delivery and enhance patient care across various medical specialties and settings.

The hypothesis proposes that VQA systems hold immense promise for revolutionizing healthcare by leveraging advanced technologies to augment human intelligence, improve diagnostic capabilities, and ultimately enhance the quality of care provided to patients.

Here are some examples illustrating the potential applications and benefits of a Visual QA (VQA) system in the healthcare industry.

- Diagnostic Assistance
- Surgical Planning
- Patient Consultations
- Medical Education
- Remote Monitoring
- Research and Development
- Clinical Decision Support

A radiologist can use a VQA system to examine the skeletal images such as X-rays, MRIs, or CT scans. By asking specific Queries concerning the presence of abnormalities, tumor characteristics, or organ functionality, the system can provide relevant insights and assist in making accurate diagnoses. Surgeons can utilize a VQA system to better understand the anatomical structures visible in pre-operative imaging studies. By asking questions about optimal incision sites, critical landmarks, or potential complications, the system can help plan surgical approaches and anticipate challenges during procedures. During patient consultations, physicians can employ a VQA system to explain medical images to patients in a more understandable manner. Patients can ask questions about their condition, treatment options, or expected outcomes, and the system can provide personalized explanations and visual aids to facilitate shared decision-making. Medical students and residents can use a VQA system as a learning tool to enhance their understanding of medical imaging and pathology. By asking questions about image interpretation, disease mechanisms, or treatment strategies, learners can receive immediate feedback and guidance to reinforce their knowledge and skills. In telemedicine settings, healthcare providers can leverage a VQA system to remotely assess patients' conditions based on uploaded images or video consultations. By asking questions about symptom severity, treatment

adherence, or recovery progress, the system can help monitor patients' health status and provide timely interventions. Researchers can employ a VQA system to analyze large-scale medical image datasets and identify patterns associated with disease progression or treatment response. By asking questions about imaging biomarkers, genetic correlations, or clinical outcomes, researchers can gain valuable insights to inform drug development, clinical trials, and precision medicine initiatives. Clinicians can integrate a VQA system into clinical decision support systems to assist in interpreting complex skeletal images and laboratory results. By asking questions about differential diagnoses, prognostic factors, or treatment guidelines, the system can provide evidence-based recommendations to guide patient care and improve outcomes. These examples highlight how a VQA system can be applied across various healthcare scenarios to enhance diagnostic accuracy, improve patient communication, support medical education, facilitate research endeavors, and ultimately, optimize the delivery of healthcare services.

1.8 Research contribution to the society

Research contributions to society for Visual QA (VQA) systems in the healthcare domain are significant and multifaceted. Some key contributions include:

- 1. Improved Diagnostic Accuracy:** VQA systems assist healthcare professionals in interpreting medical images more accurately by providing contextual information and answering specific questions about abnormalities, anatomical structures, or disease characteristics. This can lead to more exact diagnosis and treatment strategies, hence improving patient outcomes and lowering medical errors.
- 2. Enhanced Patient Care and Communication:** VQA systems facilitate better communication between healthcare providers and patients by translating complex medical information into understandable visual representations. Patients can ask questions about their condition, treatment options, or test results, and receive personalized explanations from the system, leading to increased patient satisfaction and engagement in their care.
- 3. Efficient Medical Education:** VQA systems serve as valuable Educational assets for healthcare professionals in training, residents, and practicing clinicians by offering interactive learning experiences and real-time feedback

on medical image interpretation and diagnostic reasoning. This contributes to the continuous professional development of healthcare professionals and ensures the delivery of high-quality care.

4. **Accelerated Research and Innovation:** VQA systems enable researchers to analyze large-scale medical image datasets more efficiently and extract meaningful insights into disease mechanisms, treatment responses, and prognostic factors. By automating the process of image analysis and interpretation, these systems expedite the discovery of new biomarkers, therapeutic targets, and diagnostic algorithms, leading to advancements in medical science and technology.
5. **Accessible Healthcare Services:** VQA systems have the potential to democratize access to healthcare services by extending the reach of medical expertise to underserved populations and remote areas. Telemedicine platforms equipped with Visual QA (VQA) capabilities enable patients to receive timely consultations and diagnostic assessments from specialists without requiring physical visits, thereby reducing healthcare disparities and fostering health equity.
6. **Data-driven Decision Support:** VQA systems generate valuable insights from medical imaging data that can inform clinical decision-making and guide evidence-based practice. By integrating VQA capabilities into clinical decision support systems, healthcare providers can access relevant information and recommendations at the point of care, leading to more informed treatment decisions and better patient outcomes.

Overall, research contributions in the development and implementation of VQA systems in the medical domain have the potential to revolutionize healthcare delivery, improve patient care, and advance healthcare knowledge, ultimately benefiting society as a whole.

1.9 Application of VQA

Visual Question Answering (VQA) has numerous applications across various domains, including but not limited to:

1. Healthcare:

- **Medical Image Analysis:** VQA can assist healthcare professionals in interpreting skeletal images such as X-rays, MRIs, and CT scans by answering questions about the content of the skeletal images, aiding in diagnosis and treatment planning.
- **Clinical Decision Support:** VQA systems can provide real-time assistance to healthcare providers by answering questions related to patient data, treatment protocols, and medical literature, helping them make informed decisions.
- **Patient Education:** VQA applications can be used to create interactive educational materials for patients, allowing them to ask questions about their medical conditions, treatment options, and lifestyle changes.

2. Education:

- **Interactive Learning:** VQA can enhance traditional educational materials by allowing students to ask questions about visual content such as diagrams, charts, and graphs, facilitating deeper understanding and engagement.
- **Assessment and Feedback:** VQA systems can be used to automatically generate questions based on educational content and provide immediate feedback to students, allowing for personalized learning experiences.
- **Language Learning:** VQA applications can help language learners improve their skills by providing visual prompts and answering questions about vocabulary, grammar, and cultural context.

3. E-commerce:

- **Product Recommendation:** VQA can be integrated into e-commerce platforms to assist shoppers in finding products that meet their specific

needs and preferences by answering questions about product features, compatibility, and usage.

- **Customer Support:** VQA systems can provide automated customer support by answering questions about product specifications, pricing, shipping, and returns, improving user experience and reducing the workload of customer service representatives.
- **Visual Search:** VQA can enable visual search functionality, allowing users to find products by asking questions about their appearance, brand, and category, complementing traditional text-based search methods.

4. Navigation and Assistive Technologies:

- **Smart Assistants:** VQA-powered smart assistants can assist users in navigating their surroundings, providing information about landmarks, directions, and points of interest based on visual cues captured by cameras or sensors.
- **Accessibility:** VQA applications can enhance accessibility for individuals with disabilities by providing spoken answers to questions about their environment, enabling them to interact with digital and physical spaces more independently.

5. Social Media and Content Creation:

- **Content Moderation:** VQA systems can assist social media platforms in moderating content by automatically detecting and answering questions about potentially inappropriate or harmful visual content.
- **Content Generation:** VQA can be used to generate captions, descriptions, and tags for images and videos uploaded to social media platforms, enhancing content discoverability and engagement.

These are just a few examples of the diverse applications of Visual Question Answering across different domains. As technology advances, we should expect to see more inventive applications of VQA in the future.

1.10 Organizations of Thesis

This thesis report is critical for presenting your study findings in an understandable, logical, and complete manner. The following is a potential thesis organization

structure. This was followed by a literature review, which included a review of relevant material and past research in the subject of visual question answering systems in general, as well as specific applications. The application focuses on medical care and radiology imaging, with several question and answer pairings. Chapter 2 covered four types of literature reviews: a review of the visual question answer system, a review of the radiology image dataset, research on current methodology techniques and algorithms, and finally a discussion on the strategy of extracting the features of both visual and textual datasets. Then identify the issues and gaps in the following chapters. Then, discuss the problem statement and the proposed approach's objectives. Then we'll talk about key concepts, theories, and methodology. The methodology chapter provides an overview of the research design and technique used. Explanation of data collection methods, tools, and procedures. Discuss the data analysis methodologies and statistical methods used. Presentation of study results, data, and analysis. To highlight crucial findings, use tables, figures, and graphs. Discuss any surprising findings or abnormalities. Then Discussion of how to interpret results in light of the research topic or hypothesis. Results were compared to prior literature. Analyze the study's merits, shortcomings, and implications. Finally, consider the conclusion and future scope. Summary of key discoveries and their significance. Restate the research question and aims. Recommendations for future research or applications.

1.11 Problem Statement for Visual Question Answering System

In the contemporary era, various technological applications have been developed in the healthcare domain, and answering questions related to medical images is a significant concept in the medical sector. Several technologies assist in understanding the human health status through radiographic images. In many hospitals, a plethora of scanning techniques and instruments are employed to gain insights into human health. Radiology visuals consist of scanned images, and various levels of images can be obtained based on the patient's health condition. Existing approaches for medical imaging are often inadequately supported, and the collection of medical datasets poses challenges due to their limited local availability.

Present VQA models employ CNNs to extract localized feature vectors for specific regions and LSTMs to encode feature vectors for the corresponding questions.

Nevertheless, when the response encompasses two neighboring local regions in the skeletal image and the query comprises a complex sentence, the accuracy of the attention mechanisms' answers is not entirely satisfactory. The current computer-aided diagnosis technology is generally limited to a single condition. The complexity of supplementary diagnostic technology, which utilizes analysis of a singular skeletal imaging type to deliver a thorough, specific portrayal of a patient's condition akin to a clinician's diagnosis, presents a significant constraint. The existing approach is unsuitable for dealing with such complex data, and vector construction produces inferior outcomes. The current method can only remember the preceding contextual details of the question and does not have the capability to make use of subsequent information. This limitation leads to errors in extracting the question feature.

1.12 Objectives for VQA's proposed approach

1. To propose an optimal feature extraction method for radiology images in the VQA system:

- This objective aims to identify and develop a feature extraction method specifically tailored for radiology images in the visual or imaginary QA (VQA) system.
- The model should effectively capture the relevant visual information from radiology images, such as anatomical structures, abnormalities, and other diagnostic features.
- Various techniques for feature extraction, including conventional techniques and deep neural network methodologies, will be explored and evaluated.
- The optimal feature extraction method should balance computational efficiency with the ability to represent complex visual patterns present in radiology images accurately.
- Performance metrics such as feature distinctiveness, discriminability, and computational cost will be considered in selecting the optimal method.

2. To offer an automated solution for answering the user's questions in radiology images:

- This objective involves designing and implementing an automated system that can generate accurate answers to questions asked by users based on medical images.
- The system will integrate the proposed optimal feature extraction method with advanced natural language processing (NLP) techniques for question understanding and answer generation.
- It will utilize the models based on deep learning, like Recurrent Neural Networks (RNNs) or Transformers to efficiently handle textual queries and generate suitable responses.
- The system will be trained on a high dimensional dataset of skeletal images paired with corresponding question-answer pairs to learn the associations between visual features and textual queries.
- Emphasis will be placed on the machine's accuracy, speed, and growth potential to manage a broad spectrum of medical images and questions effectively.

3. To suggest an innovative method for dictionary formation concerning images and Question-Answer (QA) pairs:

- This objective aims to develop an innovative approach for developing a comprehensive dictionary that maps visual features extracted from medical images to corresponding question-answer pairs.
- The algorithm will utilize advanced machine learning and data mining techniques to automatically identify and categorize relevant visual and textual features present in the image-QA pairs.
- It will consider semantic similarities, contextual information, and relevance scores to establish meaningful associations between images and their corresponding questions and answers.
- The dictionary creation algorithm will be designed to be scalable and adaptable, allowing it to accommodate new image-QA pairs and update existing mappings over time.
- The resulting dictionary will serve as a valuable resource for the automated VQA system, facilitating accurate and contextually relevant question answers.

4. To analyze the performance of the proposed approach in comparison to established techniques and showcase the originality of the proposed strategy:

- This objective involves conducting a comprehensive evaluation of the proposed methodology against current techniques and cutting-edge methods in the field of healthcare VQA.
- Evaluation criteria such as accuracy, precision, recall, F1-score, and computational efficiency will be used to assess the effectiveness and robustness of the proposed methodology.
- Comparative experiments will be conducted using benchmark datasets and real-world medical image-QA pairs to emphasize the strengths of the proposed method.
- Statistical analysis and qualitative assessments will be performed to highlight the strengths, limitations, and unique capabilities of the proposed methodology compared with existing approaches.
- The novelty of the proposed technique will be validated through empirical results, innovative design choices, and contributions to advancing cutting-edge methods in medical VQA research.

By achieving these objectives, the proposed research aims to advance the field of medical VQA by introducing novel methodologies, algorithms, and systems that enable accurate, automated, and contextually relevant question answering based on radiology images.

REVIEW OF LITERATURE



The use of supporting diagnostic digitally-enabled entirely on the examination of a single form of skeletal imaging has a significant constraint in producing accurate and concise descriptions of a patient's condition that are comparable to a diagnostic evaluation by a clinician. Addressing this constraint, the Healthcare Image QA model delivers on its promise by offering a feasible solution. Despite having access to a large volume of training data, the existing Healthcare Image QA datasets frequently encounter difficulties that require improvement.

The presence of faulty data within these datasets can lead to a decline in classification accuracy, even with the support of substantial training data. Therefore, refining the quality of Healthcare Image QA datasets is of utmost importance to enhance the system's overall performance and reliability.

2.1 Review of relevant literature and previous research:

- Review on Visual Question Answer System
- Review on Radiology Image Datasets
- Research on Current Methodology Techniques and Algorithms
- Feature Extraction Techniques for Visual and Textual Datasets

2.1.1 Review on Visual Question Answer

The significance of Healthcare Image QA lies in its potential impact on scientific research. By effectively combining visual question answering with medical imaging, it opens new avenues for supporting medical professionals in their diagnostic process and facilitates advancements in the field of skeletal image analysis. However, continuous efforts to address data quality and further refine the system will be crucial in leveraging the complete capacity of Healthcare Image QA in medical research and clinical applications.

The subject matter of Visual QA (VQA) for skeletal imaging is still in its early phases, with many unknown technologies and issues to handle. Because there are few standardized data sources in the medical arena, it is critical to make the Healthcare Image QA model data adaptable. This study provides numerous options to make the Healthcare Image QA system more accessible for patient consultation and medical research, laying the groundwork for future research.

The graph in Figure [2] illustrates a remarkable evolution in research output within the field of visual question answering (VQA). Prior to 2015, the volume of scholarly articles on this subject was notably sparse. The dataset for this graph is derived from a Google Scholar search query using the specified qualification of "visual question answering" and is organized by year.

Beginning in 2017, there has been a substantial surge in research activities within the domain. Over the six-year span from 2015 to 2021, the annual count of research articles has surged from 73 to 3,400—an increase of over 40 times. This exponential growth underscores a burgeoning interest and engagement in the field of VQA.

The escalating trend in scholarly contributions suggests a heightened enthusiasm and recognition of the significance of VQA within the broader academic and research community. This surge in research output also signifies a collective endeavor to address challenges, explore innovations, and advance the understanding and application of visual question answering methodologies.

Count vs. Year

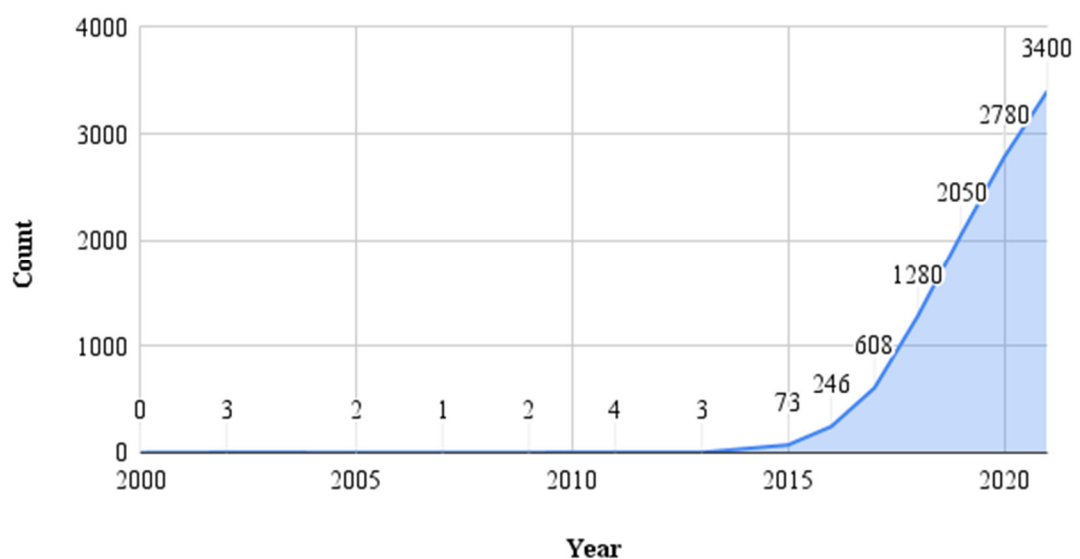


Fig. 2 : Trends in Visual Question Answering Research

The raw data regarding the number of Visual Question Answering (VQA) articles per year is sourced directly from Google Scholar. This dataset reflects the quantitative representation of scholarly publications in the field of VQA over the specified time frame. The figures demonstrate the evolving landscape of research activities and scholarly contributions within the VQA domain.

This information serves as a foundational basis for understanding the trajectory of research output, highlighting the growing interest, and providing valuable insights into the expanding body of knowledge in Visual Question Answering.

Some of the specific challenges faced by the Healthcare Image QA system include:

1. **Processing Medical-Specific Vocabulary:** Medical texts and images often contain specialized medical terminology that requires specific processing to be understood accurately by the VQA system.
2. **Combining Multi-Modal Features:** Integrating information from various sources, such as skeletal images and textual descriptions, at different levels poses a challenge that needs to be addressed effectively.
3. **Addressing Question-Visual Interaction:** Understanding the combination of the questions and the visual information derived from medical texts is crucial for accurate and contextually relevant answers.

To advance the capabilities of Healthcare Image QA, researchers need to focus on developing innovative solutions to handle these challenges and improve the model's performance. By addressing the specific needs and complexities of the medical domain, Healthcare Image QA can become a valuable tool for medical professionals, aiding in patient care and advancing medical research. However, continuous efforts and research are essential to leveraging the complete capacity of Healthcare Image QA in the medical field.

The visual transformer model employed in [1], incorporating a textual encoder transformer and a multi-modal decoder, is utilized for answer generation. The study encompasses two distinct datasets, namely PathVQA and Radiological Image Question Answering Platform, both comprising radiology images.

The study outlined in [2] focuses on the analysis and comparison of various techniques. Through the application of feature extraction methods, the analysis aims to discern prevalent faults within datasets, thereby facilitating an examination of alternative approaches to addressing Visual Question Answering (VQA) tasks. The authors underscore that, while prior endeavors have sought to mitigate linguistic bias, their approach capitalizes on the capacity to comprehend context without diminishing the significance of individual instances.

To facilitate this analysis and comparison, the study employs alternative methodologies, including the use of the Consolidated Tool for vqa. Benchmarking is incorporated, which not only evaluates model accuracy but also considers uncertainties and biases, providing valuable insights into their behavioral patterns. Additionally, interactivity is introduced, allowing end-users to select metrics for analysis and determine the scope of data evaluation. The investigation into Multimodal continuous visual Attention mechanisms underscores the potential drawback of discrete attention mechanisms—despite their exceptional versatility, there exists a risk of losing focus due to the generation of scattered attention maps. Various methodologies are currently under examination, including "Unshuffling Data for Improved Generalization in Visual Question Answering," "Structured Multimodal Attentions for TextVQA," and "Zero-shot Visual Question Answering Using Knowledge Graph," among others.

The computational process known as VQA involves the input of an image and a corresponding question, with the computer generating the correct answer to the query. The aspiration within the realm of AI research has long been the development of robots capable of comprehending visual information and providing responses akin to human understanding. Notably, recent recognition has been accorded to research endeavors in the domain of VQA. Specifically, in the context of medical visual question answering (Med-VQA), a clinical inquiry is paired with a radiological image. [3]

The work by [4] provides an in-depth analysis of methodologies, findings, potential advancements, and challenges in the field. The paper delves into contemporary datasets sourced from reputable outlets such as journals, conferences, and pertinent articles, with a specific focus on computational multimedia in skeletal image computing and computer-assisted intervention. It meticulously delineates the four integral components of the framework, namely the Image representation module, Language representation module, Multi-modal integration unit, and Answer generation component. This comprehensive examination contributes valuable insights to the scholarly discourse in the domain.

The Review of attention method used in [5] where multimodal fusion technique is used for both visual and textual feature extraction also discussed the classification application of existing attention mechanisms.

The review paper serves as a comprehensive enhancement on the notable advancements in visual question answering (VQA) utilizing images, particularly focusing on recent developments. Drawing insights from the referenced study [7], the review underscores the growing importance of multimodal approaches in enhancing visual question answering systems.

The exploration of various aspects and benefits of visual question answering is a prominent feature of this review. It builds upon the foundation laid by the referenced study and incorporates subsequent updates in the field, offering a nuanced and up-to-date perspective on the subject matter. The formal tone maintains the academic rigor appropriate for a review of this nature.

A limited number of surveys have delved into the realm of Visual Question Answering (VQA), addressing diverse methodologies for accomplishing this task and the introduction of new datasets to enhance existing benchmarks. Existing surveys predominantly aim to establish an organizational framework for the models and datasets employed in VQA, with some concentrating on specific subdomains of this field [4], while others present a more expansive exploration of the subject [8, 7, 9, 10]. This section offers a fundamental comparison between the present work and prior surveys within the domain. The tone maintains a formal demeanor suitable for scholarly discourse.

The study outlined in [8] offers a comprehensive introduction to this research domain, encompassing classical and established Visual Question Answering (VQA) datasets alongside emerging ones. The paper delves into evaluation metrics, providing insights into understanding and gauging various aspects of VQA models. Additionally, it explores diverse architectural approaches utilized in VQA, considering aspects like scene-text incorporated into specific datasets [32, 33]. On the other hand, [7] embarks on a meticulous evaluation in VQA, furnishing intricate descriptions and explanations concerning current methodologies, datasets, and evaluation procedures. The study critically assesses the present landscape of the field and contemplates potential future

trajectories. The tone adheres to a formal style suitable for academic discourse. The work delineated in [9] encompasses the recent strides in Visual Question Generation (VQG), a pivotal facet of Visual Question Answering (VQA), focusing on the creation of new datasets. The authors scrutinize the prevailing techniques in VQG, methodologies for evaluating the efficacy of generated questions, prevalent algorithms in this subfield, and the extant challenges. On a parallel note, [9] furnishes a comprehensive exploration of the tools and approaches employed for answering queries related to skeletal imaging. This survey meticulously delves into notified datasets within the healthcare domain, evaluating approximately 45 papers. Furthermore, [10] scrutinizes existing VQA datasets, metrics, and models, providing a comprehensive assessment of their advancements and persisting challenges. The tone maintains a formal style suited for scholarly communication.

2.1.2 Review on Radiology Image Datasets

Establishing a robust dataset for Visual Question Answering (VQA) poses a formidable challenge, necessitating the involvement of numerous annotators and domain experts. This is particularly crucial in the context of deep learning, where substantial amounts of data are essential for models to generalize effectively. The process is intricate, and not all datasets are constructed *de novo*; some opt for the creation of new subsets within existing inputs. In this section, we delve into an exploration of the frequently employed datasets derived from the scrutinized publications, shedding light on their significance in advancing research in the field of VQA. This meticulous approach ensures a comprehensive understanding of the datasets that underpin the advancements in VQA.

There are various public-available skeletal images VQA datasets up to date Healthcare Image QA-2018 [67], Radiological Image QuestionAnswering Platform [68], Healthcare Visual Q&A 2019 [69], RadVisDial [70], PathVQA [71], Healthcare Image QA-2020 [72], SLAKE [74], and Healthcare Image QA-2021 [73]. Additional dataset is used for research such as VQA v1, VQA v2, VQA CP v1, VQA CP v2.

Healthcare Image QA-2018 [67] represents a landmark dataset introduced through ImageCLEF 2018, marking the inception of publicly available datasets in the medical domain. The dataset's creation employed a semi-automatic methodology for generating Question-Answer (QA) pairs based on image captions. A rule-based

Question Generation (QG) system played a central role, simplifying sentences, identifying answer phrases, creating questions, and subsequently ranking candidates. The QA pairs produced by this system underwent meticulous evaluation, being scrutinized twice by professional human annotators, one of whom possessed expertise in clinical medicine. This dual-check process involved validating semantic consistency and ensuring clinical relevance in relation to the associated medical images. This approach not only laid the foundation for Healthcare Image QA-2018 but also set a benchmark for the careful curation and validation of datasets in the medical visual question answering domain.

Radiological Image Question Answering Platform [68], unveiled in 2018, is a dataset crafted specifically for radiology applications. The dataset's image collection exhibits balance, featuring examples from the head, chest, and abdomen sourced from MedPix. In a bid to simulate a real-world scenario, the author presented these images to medical professionals, prompting them to pose spontaneous inquiries. Clinicians were tasked with formulating questions in both free-form and template formats. Subsequently, the generated Question-Answer (QA) pairs underwent manual scrutiny and categorization to ascertain their clinical focus. The answers in this dataset encompass both closed and open-ended formats. Despite its modest size, the Radiological Image Question Answering Platform dataset furnishes valuable insights for the development of AI systems tailored for radiological applications.

Healthcare Visual Q&A 2019 [69] constitutes the second iteration of the Healthcare Image QA series, introduced as part of the CLEF Image Retrieval and Classification Task 2019 challenge. Aligned with the Radiological Image Question Answering Platform [68] model, Healthcare Visual Q&A 2019 specifically targets four prevalent question categories: modality, plane, organ system, and abnormality. These question classifications were derived from patterns identified in hundreds of spontaneously generated and validated questions sourced from Radiological Image Question Answering Platform [68]. While the first three categories (modality, plane, and organ system) are amenable to classification problem-solving approaches, the fourth category (abnormality) introduces a more intricate challenge, necessitating answer generation capabilities. The outline of the healthcare VQA datasets and their fundamental qualities are described. VQA 2.0 [75] has 204000 images and 614000

question answer pair datasets. It is very huge to load it into a model. The source of these datasets is Microsoft COCO [76]. Question Answers are created manually and the categorization of questions in different types, like object, color, sport, count, etc. The Healthcare Image QA 2018 [67] dataset has 2,866 image data points and 6413 question answer pairs. The source of the image and content is Pub Central Articles. The creation of questions and answers is synthetical. The question categories are mentioned, such as location, finding, Yes/ No questions, and other questions. From Radiological Image QuestionAnswering Platform [68], there are 315 image and 3515 question answer pairs present; the source of images and content is the MedPix database contains head axial single-slice CTs or MRIs, chest X-rays, and abdominal axial CTs. The question answer creation is in Natural; the question categories are Modality, Plane, Organ System, Abnormality, Object/Condition Presence, Positional Reasoning, Color, Size, Other Attributes, and Counting.

In Healthcare Visual Q&A 2019 [69], 4,200 images were used and 15,292 question answer pairs were created. The source of images and content is the MedPix database, which is various in 36 modalities, 16 planes, and 10 organ systems. The question answer creation is synthetic; here the question categories are modality, plane, organ system, and abnormality. RadVisDial [70] for Silver-standard has 91,060 images and 455,300 question answer pairs. The source of the image and content is MIMIC_CXR [77] Chest X-ray posterior-anterior (PA) view. The creation of the question answer is Synthetical and the question category is abnormality. RadVisDial [70] for Gold-standard has 100 images and 500 question answer pairs. The source of the image and content is MIMIC_CXR [77]. Chest X-ray posterior-anterior (PA) view. The creation of the question answer is natural, and the question category is Abnormality. In PathVQA [71], 4,998 images and 32,799 question answer pairs were used, and the source of the images and content is electronic pathology textbooks in the PEIR Digital Library. The creation of questions and answers is synthetic. The categories of question from PathVQA [71] are color, location, appearance, shape, etc.

Healthcare Image QA 2020 and Healthcare Image QA 2021 [72, 73] used 5,000 images and 5,000 question answer pairs. The source of the image and content is from MedPix databases, the question was synthetical, the category of the question is abnormality. The dataset SLAKE [78] has 642 images and 14000 question answer

pairs. The source of the images and content is medical segmentation decathlon [79], NIH chest X-ray [80], and CHAOS [81] with chest X-rays/CTs, abdomen CTs/MRIs, head CTs/MRIs, neck CTs, and pelvic cavity CTs. The creation of the question method is natural, and the categories are organ, position, knowledge graph, abnormality, modality, plane, quality, color, size, and shape.

Table 1 : Descriptive Statistics of Existing visual and textual dataset

Image Dataset		Question Answer Pair Dataset	
Mean	31818.1	Mean	115181.9
Standard Error	21051.62017	Standard Error	70964.36826
Median	4599	Median	10206.5
Mode	5000	Mode	5000
Standard Deviation	66571.06818	Standard Deviation	224409.0364
Sample Variance	4431707119	Sample Variance	50359415620
Kurtosis	5.629002365	Kurtosis	2.227660663
Skewness	2.408398774	Skewness	1.892712375
Range	203900	Range	613500
Minimum	100	Minimum	500
Maximum	204000	Maximum	614000
Sum	318181	Sum	1151819
Count	10	Count	10
Largest(1)	204000	Largest(1)	614000
Smallest(1)	100	Smallest(1)	500
Confidence Level (95%)	47622.07327	Confidence Level (95%)	160532.5536

Table 2 : Correlation of Existing visual and textual dataset

	Image dataset	Question Answer Pair dataset
Image	1	
Question Answer Pair	0.9696640705	1

Table 3 : Covariance of Existing visual and textual dataset

	Image dataset	Question Answer Pair dataset
Image	3988536407	
Question Answer Pair	13037360656	45323474058

Table 4 : Cumulative Frequency of Existing visual and textual dataset

Image Dataset			Question Answer Pair Dataset		
Bin	Frequency	Cumulative %	Bin	Frequency	Cumulative %
100	1	5.56%	5000	4	22.22%
315	1	11.11%	100	1	27.78%
500	1	16.67%	315	1	33.33%
642	1	22.22%	500	1	38.89%
2866	1	27.78%	642	1	44.44%
3515	1	33.33%	2866	1	50.00%
4200	1	38.89%	3515	1	55.56%
4998	1	44.44%	4200	1	61.11%
5000	4	66.67%	4998	1	66.67%
5000	0	66.67%	6413	1	72.22%
5000	0	66.67%	14000	1	77.78%
5000	0	66.67%	15292	1	83.33%
6413	1	72.22%	32799	1	88.89%
14000	1	77.78%	91060	1	94.44%
15292	1	83.33%	455300	1	100.00%
32799	1	88.89%	5000	0	100.00%
91060	1	94.44%	5000	0	100.00%
455300	1	100.00%	5000	0	100.00%
More	0	100.00%	More	0	100.00%

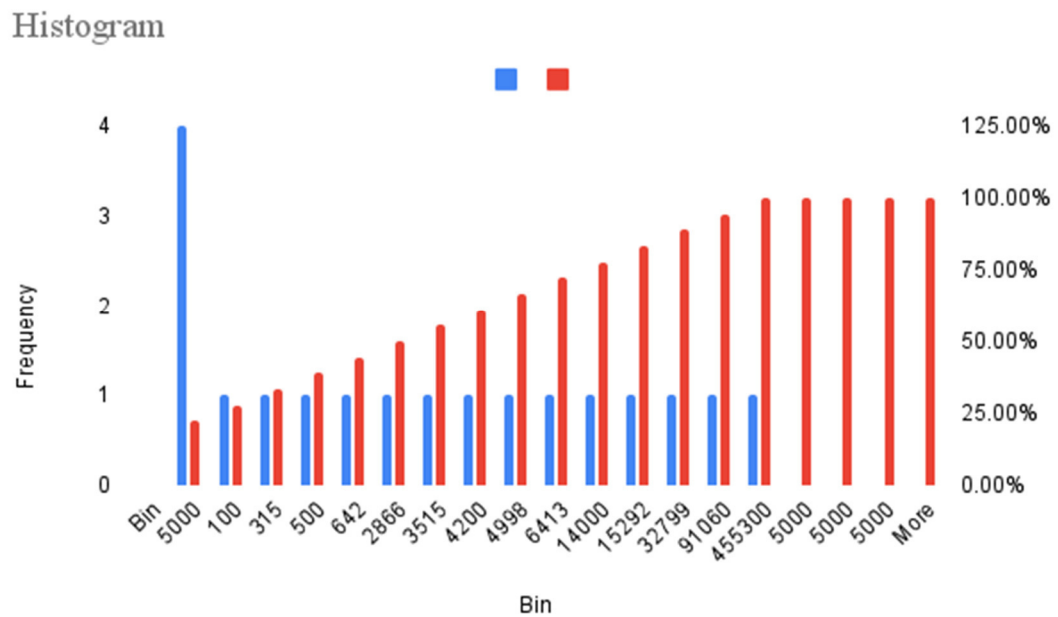


Fig 3 : Histogram of existing collective dataset with its frequency

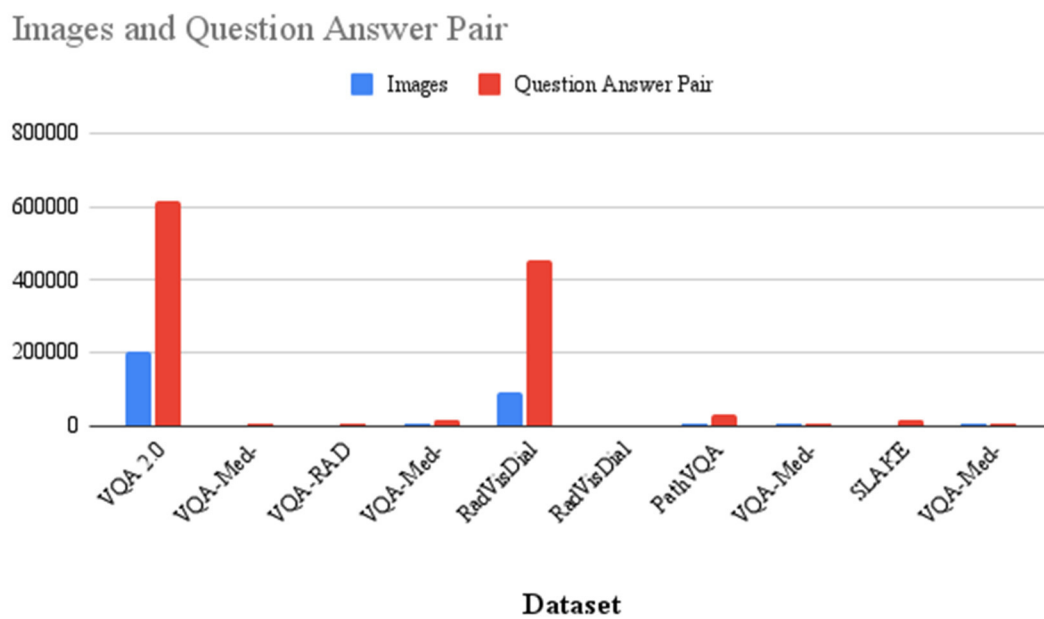


Fig 4 : Total data present in various image and question answer pair dataset

2.1.3 Research on Current Methodology with existing techniques and Algorithms

In their research, Fuji Ren et al. [11] introduced an innovative model named CGMVQA, which integrates categorization and answer generation skills to dissect the intricate Visual Question Answering (VQA) task into smaller components. The model involves tokenizing text and incorporating image data, employing the pretrained ResNet152 architecture to consolidate three disparate varieties of embeddings and extract visual features for handling textual data.

The CGMVQA model performed exceptionally well on the CLEF Image Retrieval and Classification Task 2019 Healthcare Image QA dataset, with a rate of classification of 0.6401, a word matching rate of 0.658, and a conceptual similarity score of 0.679. These results demonstrated the model's superiority over existing approaches to accurately answering medical visual questions.

The study suggested that CGMVQA holds the potential to assist medical professionals in clinical examination and diagnosis, providing valuable insights and support in the medical field. By effectively addressing medical visual questions, the CGMVQA model can aid medical professionals in making informed decisions and improving patient care.

Lubna A. et al.'s investigation into Visual QA (VQA) for medical images within the context of the CLEF Image Retrieval and Classification Task 2019 medical VQA dataset [12] focused on addressing the intricacies associated with answering questions tailored to different medical image modalities. These modalities encompass X-rays, computed tomography (CT), ultrasound (US), magnetic resonance imaging (MRI), and various others. This study's approach consisted of two steps: The input medical image is first classified into its respective modality class using a convolutional neural network (CNN), and the correct solution for the VQA problem is then delivered based on the CNN output. The model achieved a testing accuracy of 83.8%, demonstrating its effectiveness in answering questions related to diverse medical image modalities.

This performance was equivalent to the methods used during that time, showcasing the potential of the proposed approach in addressing the challenges of VQA in the medical domain.

In their research, Fazal Muhammad et al. [13] utilized reverse frequency allocation (RFA) and emphasized both decoupled DL-UL association (De-DUA) and coupled DL-UL association (Co-DUA) techniques to explore their performance in wireless communication systems.

In the De-DUA approach, a user was randomly connected to base stations (BSs) belonging to two separate tiers—one for downlink (DL) communication and another for uplink (UL) communication. This decoupling of DL and UL associations aimed to enhance the system's performance.

Additionally, the researchers employed reverse frequency allocation (RFA) as part of their methodology. RFA involves allocating frequencies in the reverse direction compared to conventional frequency allocation schemes, potentially improving overall system performance.

The study's findings indicated that De-DUA, combined with RFA, outperformed Co-DUA with regard to coverage performance. This suggests that the Decoupled DL-UL Association, along with the utilization of Reverse Frequency Allocation, offers advantages over the traditional Coupled DL-UL Association in the context of wireless communication systems.

In their research, Dhruv Sharma et al. [14] developed MedFuseNet, an attention-based multimodal deep learning network specifically designed for visual question answering (VQA) on medical images. MedFuseNet aimed to address the complexities involved in VQA for medical images by decomposing the problem into simpler components and maximizing learning efficiency.

MedFuseNet utilized attention mechanisms, directing the model's attention towards the important regions of the medical images while processing the associated questions. This attention-based approach aimed to amplify the model's effectiveness and interpretability.

The study addressed two critical aspects of VQA for medical images: (1) categorization of the images and (2) generation of two different types of answer predictions. By addressing these aspects, MedFuseNet provided more comprehensive and accurate answers to the given medical questions.

The experimental results demonstrated that MedFuseNet outperformed state-of-the-art VQA techniques for skeletal images, showcasing its superior performance in answering medical visual questions. Additionally, the attention visualization provided illuminating the model's decision-making process, increasing its interpretability and transparency.

The development of MedFuseNet highlights the potential of attention-based multimodal deep learning networks in advancing VQA systems for skeletal images, offering valuable support to medical professionals in clinical decision-making and enhancing the understanding of the model's predictions.

In their research, Shengyan Liu et al. [15] introduced a novel bi-branched model named BPI-MVQA, which stands for Parallel Networks and Image Retrieval for Medical Visual QA. The BPI-MVQA model was designed specifically for Medical Visual Question Answering (MVQA) tasks.

The initial branch of the BPI-MVQA model utilized a transformer topology based on a parallel network, allowing for the effective extraction of both image sequence features and spatial features, providing complementary benefits for the MVQA task.

To fuse the multi-modal features extracted from medical images and textual questions, the researchers employed a multi-head self-attention mechanism. This method allowed for the implicit fusion of information from different modalities, enhancing the model's ability to process diverse information sources.

The second branch of the BPI-MVQA model used the relative proximity of image features provided by the VGG16 network to produce suitable text labels. This approach allowed for effective image retrieval and the association of the most relevant text labels with the given medical images.

Experimental results demonstrated that the BPI-MVQA model achieved state-of-the-art performance on three Healthcare Image QA datasets. The model's cutting-edge results showcased its effectiveness in answering medical visual questions, making it a valuable tool in medical image analysis and diagnosis.

In this study, the researchers introduced a new and innovative bi-branched model called BPI-MVQA, aiming to tackle medical visual question-answering tasks. The

BPI-MVQA model utilizes parallel networks and image retrieval techniques to enhance its performance.

The first branch of BPI-MVQA incorporates a transformer structure based on a parallel network. This design enables the model to extract spatial and image sequence features efficiently, benefiting from the complementary strengths of both approaches.

To merge the multi-modal characteristics effectively, the researchers employed a multi-head self-attention mechanism. This mechanism allows the model to implicitly combine information from various modalities, enhancing its ability to process diverse visual and textual information.

The second branch of the BPI-MVQA model leverages the visual features collected by the VGG16 network. These visual features are then used to generate relevant text labels, aiding in the association of appropriate textual information with the given medical images.

The proposed BPI-MVQA model represents a promising advancement in medical visual question answering, as it combines parallel networks, transformer structures, multi-head self-attention, and image retrieval techniques to accomplish top-tier performance in responding to medical visual queries.

Rahhal and Mohamad Mahmoud AI [16] proposed a method for extracting visual information using the Vision Transformer (ViT) paradigm and a transformer encoder-decoder structure. The system generates autoregressive answers by combining textual and visual representations and using a multi-modal decoder. The proposed model was verified against radiological imaging datasets from the Radiological Image QuestionAnswering Platform and PathVQA.

Visual Question Answering (VQA) is a recent advancement in computer vision aimed at improving picture captioning by allowing users to ask questions about specific characteristics of images [17]. Transformers, unlike recurrent neural networks (RNNs), learn connections between sequence components rather than processing them recursively and considering only the current context. Transformer designs facilitate long-range associations by attending to entire sequences.

One of the frequently used models for representing textual data is BERT (Bidirectional Encoder Representations from Transformers) [18]. BERT is a language

model that uses large-scale unsupervised corpora and a bidirectional attention mechanism to generate context-sensitive representations for each word in a given phrase.

By incorporating the Vision Transformer model and leveraging the power of transformers like BERT, Rahhal's proposed method demonstrates promising potential for addressing visual question answering tasks and enhancing the understanding of visual content in medical imaging datasets.

To extract comprehensive image features, we propose using a parallel structure based on ResNet152 [19, 20] and Gate Recurrent Unit (GRU) [21]. This approach allows us to capture both full-scale image features and local features effectively.

For preserving spatial feature data from images captured in various dimensions, we retain sequential encoding of the feature information from the original three-channel images. Subsequently, we convert these images into single-channel grayscale images and pass them through the stacked GRU network.

The characteristics received from the GRU network, as well as the features created by each layer of ResNet152, are then merged to provide comprehensive and informative image features. This combination of characteristics from the ResNet152 and GRU networks ensures that the visual content is fully represented.

By adopting this parallel structure approach, we can effectively capture both global and local image features, enabling better image understanding and enhancing the performance of visual analysis tasks.

The main building block of our multi-classification model is the transformer structure, which has proven to be effective in understanding complex biomedical literature. To achieve this, we leverage the power of Biobert [22], which surpasses the performance of Bidirectional Encoder Representations from Transformers (Bert) [23] in various biomedical text mining tasks and biomedical data training.

In contrast to the traditional input format of the Bert model, we adopt a novel approach. Instead of using just the textual information, we concatenate both the picture features and question features as the input to the transformer. By leveraging the diverse qualities of both types of features, we aim to improve the model's understanding of the data.

To further enhance the model's performance, we introduce the multi-head self-attention process. This innovation allows the model to effectively integrate and process the input properties, leading to better outcomes in our multi-classification tasks.

By synthesizing the strong points of the transformer structure, BioBERT, and the multi-head self-attention mechanism, our model demonstrates promising potential in biomedical text mining and classification tasks. It allows for a comprehensive understanding of complex biomedical data, contributing to advancements in the biomedical field.

The growth of visual question answering in the medical arena (Healthcare Image QA) has resulted in the birth of various novel ways for achieving VQA goals. These strategies may also be useful in the field of Healthcare Image Quality Assurance. In healthcare image quality assurance, the feature extractor is conventionally a traditional convolutional neural network (CNN) that has been pre-trained using ImageNet. On the other hand, the picture feature extractor often uses a recurrent neural network (RNN) or a transformer-based model.

One specific model that has been proposed for Healthcare Image QA is the multi-modal factorized bilinear pooling model (MFB) by Peng et al. [24]. This model is a deep network that combines ResNet152 and LSTM (long Short-Term memory) components. The MFB model is designed to effectively pool information from different modalities and encode it into a unified feature representation, enhancing the model's ability to answer questions based on both visual and textual information.

The integration of these innovative methods, such as MFB, with classical CNNs and RNNs can lead to more advanced and powerful Healthcare Image QA systems. These models open up new possibilities for medical professionals in clinical analysis, diagnosis, and research by providing accurate and interpretable answers to skeletal visual questions.

In the CLEF Image Retrieval and Classification Task 2019 Healthcare Image QA competition, the Zhejiang University team secured first place with their creative model [25]. Their model incorporated Bert to extract question characteristics and visual attributes from the middle layer of VGG16. This innovative combination

allowed for effective information extraction from both textual and visual inputs, leading to their success in the competition.

Kornuta et al. [26] The study introduced a modular pipeline architecture grounded in transfer learning and multitask learning methodologies. This approach enabled them to achieve impressive results in the ImageCLEF competition, showcasing the power of combining different learning techniques in a structured manner.

For the ImageCLEF2021 Healthcare Image QA test, Liao et al. [27] utilized the Skeleton-based Sentence Mapping (SSM) knowledge inference methodology. This approach allowed them to infer relevant knowledge from the textual information, contributing to their success in the competition.

Al-Sadi et al. [28] finished second in the ImageCLEF 2021 Healthcare Image QA exam by efficiently using data augmentation approaches. Data augmentation is the process of creating new training data by performing modifications on existing data, which improves the model's robustness and performance.

To address various difficulties in Med-VQA, Zhang et al. [29] proposed a novel conditional reasoning framework. This framework automatically develops suitable reasoning techniques for different Med-VQA challenges, showcasing the potential of adaptive reasoning in medical visual question-answering tasks.

Overall, these innovative approaches and successful models have demonstrated the potential of advanced techniques in Healthcare Image QA competitions, contributing to advancements in the fields of medical image analysis and question answering.

2.1.4 Survey on visual and textual Feature Extraction Technique

This survey investigates contemporary methodologies in feature extraction applied to multimodal datasets, encompassing both visual and textual domains. Feature extraction serves as a pivotal step in artificial intelligence systems, particularly in tasks involving computer vision and natural language understanding. The survey delves into the diverse techniques employed for extracting informative features from visual and textual data, highlighting their applications, strengths, and challenges.

There are various visual feature extraction techniques such as Traditional Computer Vision Approaches, Convolutional Neural Networks (CNNs), Transfer Learning Strategies, Attention Mechanisms in Visual Feature Extraction and ect. Bag-of-Words

Models, Word Embeddings (e.g., Word2Vec, GloVe), Transformer-Based Architectures (e.g., BERT, GPT) and Hybrid Models are the Textual Feature Extraction model.

The object detection methodology founded on deep neural networks represents a pioneering technique that has undergone substantial advancements in recent years. This method demonstrates the capacity to derive abstract high-level features by amalgamating low-level features from samples. The resulting characteristics exhibit robust expressive and generalization capabilities, marking a significant stride in the evolution of computer vision methodologies. The object detection technique based on candidate boxes, commonly referred to as a two-stage algorithm, follows a distinctive process involving region proposal extraction and subsequent candidate box recognition and regression. An example in this category is the R-CNN series [82, 83, 84]. R-CNN [84] initiates the process by employing a selective search method to identify candidate frames, proceeds to extract features using deep neural networks, and concludes with support vector machines for target classification. In the evolution of this approach, Fast R-CNN [82] has been introduced, which streamlines the process by pooling features for each candidate frame and replacing the support vector machine with a softmax classifier. A notable efficiency enhancement is achieved by extracting image features only once, contributing to accelerated training and inference speeds. Faster R-CNN [83] revolutionizes the object detection landscape by utilizing neural networks to generate candidate boxes, eliminating the need for selective search techniques. This ensures a genuinely end-to-end process for object identification. Notably, Faster R-CNN integrates convolution characteristics for region proposal, classification, and regression, fostering improved accuracy and processing efficiency. In contrast, the regression-based object detection method, exemplified by YOLO [85] and SSD [86], adopts a single-stage paradigm. This approach skips the traditional candidate box extraction stage and treats object detection as a regression problem. Neural networks are employed to determine both the categories and positions of targets in each image block, marking a departure from the two-stage algorithms.

Convolutional Neural Networks (CNNs) are a Subset of deep neural networks particularly designed for tasks involving visual data, such as image recognition, classification, and segmentation. They have become a cornerstone in computer vision

and image processing. Here are key aspects of CNNs: CNNs scan input data using convolutional layers and learnable filters or kernels.. These filters detect patterns, edges, and textures in the input. Convolutional operations help capture spatial hierarchies of features. CNNs use convolutional layers to scan the input data with learnable filters or kernels. These filters detect patterns, edges, and textures in the input. Convolutional operations help capture spatial hierarchies of features. Following convolution, pooling layers minimize the spatial dimensionality of the resulting feature maps. Max pooling is a typical approach that retains the maximum value in a set of neighboring pixels, thus downsampling the data. The convolution and pooling, fully connected layers are employed for high-level reasoning. These layers establish connections between each neuron in one layer and every neuron in the subsequent layer, creating a dense representation. Non-linear activation functions, such as ReLU (Rectified Linear Unit), are commonly utilized after convolutional and fully connected layers. ReLU adds nonlinearity to the system, facilitating its acquisition of intricate patterns. CNNs are trained by backpropagation and optimization techniques such as stochastic gradient descent. The network learns how to modify the weights of filters and neurons to reduce the discrepancy between expected and real outputs. Dropout is a regularization approach that prevents overfitting in CNNs. During training, it randomly removes a subset of neurons, driving the network to acquire more robust characteristics. CNNs that have been pre-trained on huge datasets (such as ImageNet) can be fine-tuned to do specific tasks. This transfer learning technique uses knowledge obtained from one job to boost performance on another. CNNs are capable of not only classifying images but also localizing and detecting objects within images. Techniques like region-based CNNs (R-CNN) and its variants have been successful in object detection.

The architecture of a **Deep Belief Network (DBN)** is structured as a stack of layers, including visible and hidden layers. A typical DBN comprises multiple layers of latent variables that form a hierarchical, generative model. The bottom layer of the network represents the visible layer. Nodes in this layer correspond to the observed variables or input features. This layer is when external information enters the network. There is at least one hidden layer above the visible layer. Each hidden layer captures more abstract and complicated features. The number of neurons in each hidden layer

is predetermined based on the complexity of the task at hand. Every neuron in one layer is connected to every neuron in the subsequent layer, with weights determined during the training process. To train a Deep Belief Network (DBN), each pair of adjacent layers undergoes training as a Restricted Boltzmann Machine (RBM). An RBM comprises two layers: visible and hidden, with connections between nodes within each layer but not between layers. The network is trained incrementally through unsupervised learning, with RBMs learning to reconstruct their inputs. Once trained, the DBN functions as a generative model. It can produce new samples that are similar to the training data. Following pre-training, the network can be fine-tuned with supervised learning for specific tasks such as classification or regression. Each node in the hidden layers typically uses a sigmoid activation function, facilitating the modeling of complex, non-linear relationships. The top layer of the network is frequently utilized for the task at hand. For example, in a classification task, this layer could represent the class labels. The weights between nodes are modified during training to reduce the difference between the input and the reconstructed input. The layer-wise training approach, starting from the visible layer and moving upward through the hidden layers, helps in the efficient learning of hierarchical representations of the input data. The learned hierarchical features make DBNs effective in capturing intricate patterns in data, particularly in unsupervised or generative modeling tasks.

BERT, initially designed for tasks in natural language processing, has been repurposed for multimodal applications, such as Visual Question Answering (VQA). In the context of VQA, BERT extends its capabilities to jointly understand textual and visual information. BERT is a transformer-based language model renowned for its ability to leverage bidirectional context in discerning the meanings of words within a given phrase. Pretrained on extensive corpora, BERT excels in acquiring contextual representations of words, thereby adeptly capturing intricate linguistic nuances. In VQA, BERT is used to fuse information from both the textual question and the visual content (image). The textual question is encoded using BERT's language model. The image features are typically extracted using a Convolutional Neural Network (CNN). BERT generates embeddings for the textual question, capturing its contextual information. The image features are transformed into a compatible embedding space.

The embeddings from the textual question and visual features are combined either through simple concatenation or attention mechanisms. Concatenation results in a unified representation that captures both textual and visual information. The combined representation is then fed into a classification head to forecast the answer to the given question as text. The classification head is typically a fully connected layer or a sequence of layers for answer prediction. The model is often trained in a supervised manner using datasets where each question is paired with its corresponding image and answer. The training consists of minimizing a loss function, such as cross-entropy loss, between the expected and ground truth answers. BERT for VQA can be trained on enormous amounts of linguistic data before being fine-tuned on VQA-specific datasets. Fine-tuning adapts the model to the specifics of the VQA task. BERT's bidirectional context understanding is advantageous in capturing nuanced relationships between words in questions. The multimodal fusion allows the model to leverage both textual and visual information for accurate answers. Integrating vision and language models can be computationally intensive. Managing long sequences (question + image features) might require strategic attention. BERT for VQA finds applications in various domains, including medical image analysis, robotics, and accessibility technologies. BERT for VQA leverages the strengths of transformer-based language models to jointly understand textual and visual content, enabling more context-aware and accurate answers to questions about images.

Table 5 : Describes the publications obtained and reviewed during the current survey

S.No.	Survey on existing methodology in brief highlights	Reference
1.	Bias has a negative impact on the content branch, whereas it has a beneficial impact on the context branch. This architectural innovation aims to mitigate the impact of bias within the learning process, acknowledging the nuanced relationship between content and context in order to enhance overall model performance and fairness	34
2.	Novel digital framework for analyzing, evaluating, and testing models and datasets.	35

S.No.	Survey on existing methodology in brief highlights	Reference
3.	The attention mechanisms utilized in VQA models are depicted as multimodal feature functions, aiming to emulate human attention more effectively.	36
4.	This paper proposes a new dataset and evaluation method for assessing models' generalization abilities outside of distribution.	37
5.	Enact data partitioning and training environments to mitigate false correlations while preserving genuine correlations.	38
6.	Novel modules have been introduced to facilitate the extraction of textual information from images and annotations for the TextVQA dataset. These modules represent a significant advancement in the capability to accurately read and comprehend text embedded within visual content. The incorporation of these enhancements reflects a commitment to refining the performance and versatility of systems operating on the TextVQA dataset, ultimately contributing to the progress of text-based visual question answering. The introduction of such modules aligns with the formal evolution of methodologies in the pursuit of more effective and comprehensive solutions for handling textual information within visual contexts.	39
7.	The image or visual QA algorithms are zero-shot modeled using external knowledge graphs and a new dataset.	40
8.	Creates a question-based reasoning module for healthcare Visual QA systems.	41
9.	Answers questions using a model-agnostic implication generator.	42
10.	Creates a new method for evaluating VQA models based on "skills and concepts" shown in the image.	43
11.	This technique aims to identify and mitigate negative bias	44

S.No.	Survey on existing methodology in brief highlights	Reference
	during training.	
12.	Proposes a network structure for decomposing visual concepts to provide better contextualized answers.	45
13.	A psycholinguistic approach to comprehending and addressing Visual Question Answering (VQA) and catastrophic forgetting is employed.	46
14.	A trilinear model was developed to accommodate images, questions, and information in the responses.	47
15.	Unsupervised radiological image learning using contrast and posterior representation distillation in a VQA context.	48
16.	Uses a basic yet effective bimodal fusion strategy for CQA.	49
17.	It describes an embedding method for obtaining region-of-interest and Prognosticate information from image-question pairings.	50
18.	This proposal includes a unique job for automatic image caption generation based on scene text, two datasets, and a model for solving the problem.	52
19.	This document details their participation in the 2021 Healthcare Image QA competition and the model they built.	51
20.	Proposes regularizing attention layers to enhance visual information extraction.	53
21.	Proposes adding detailed synthetic annotations to the CLEVR dataset.	54
22.	Proposes a method for training VQA models separately by mixing each trained model.	55
23.	To reduce bias in model learning, overfit biassed data and fine-tune on unbiased data.	56
24.	To compensate for the loss caused by biased functions, an objective function is created based on the language of the inquiry..	57

S.No.	Survey on existing methodology in brief highlights	Reference
25.	This concept proposes employing a dual-encoder dense retrieval to enhance VQA models with external unstructured data.	58
26.	The authors employ scanned documents and metamorphs to solve ST-VQA.	59
27.	A transformer-based model is proficient in answering questions in multiple languages.	60
28.	An interactive visual analytics tool has been developed to facilitate visual and language reasoning within transformer models.	51
29.	Transfer learning from a perfect-sighted model improves the resilience of VQA models.	61
30.	Improved visual feature extraction to enhance Visual QA performance with transformers.Increased the robustness of Visual QA models using transfer learning of a perfect-sighted model.	62

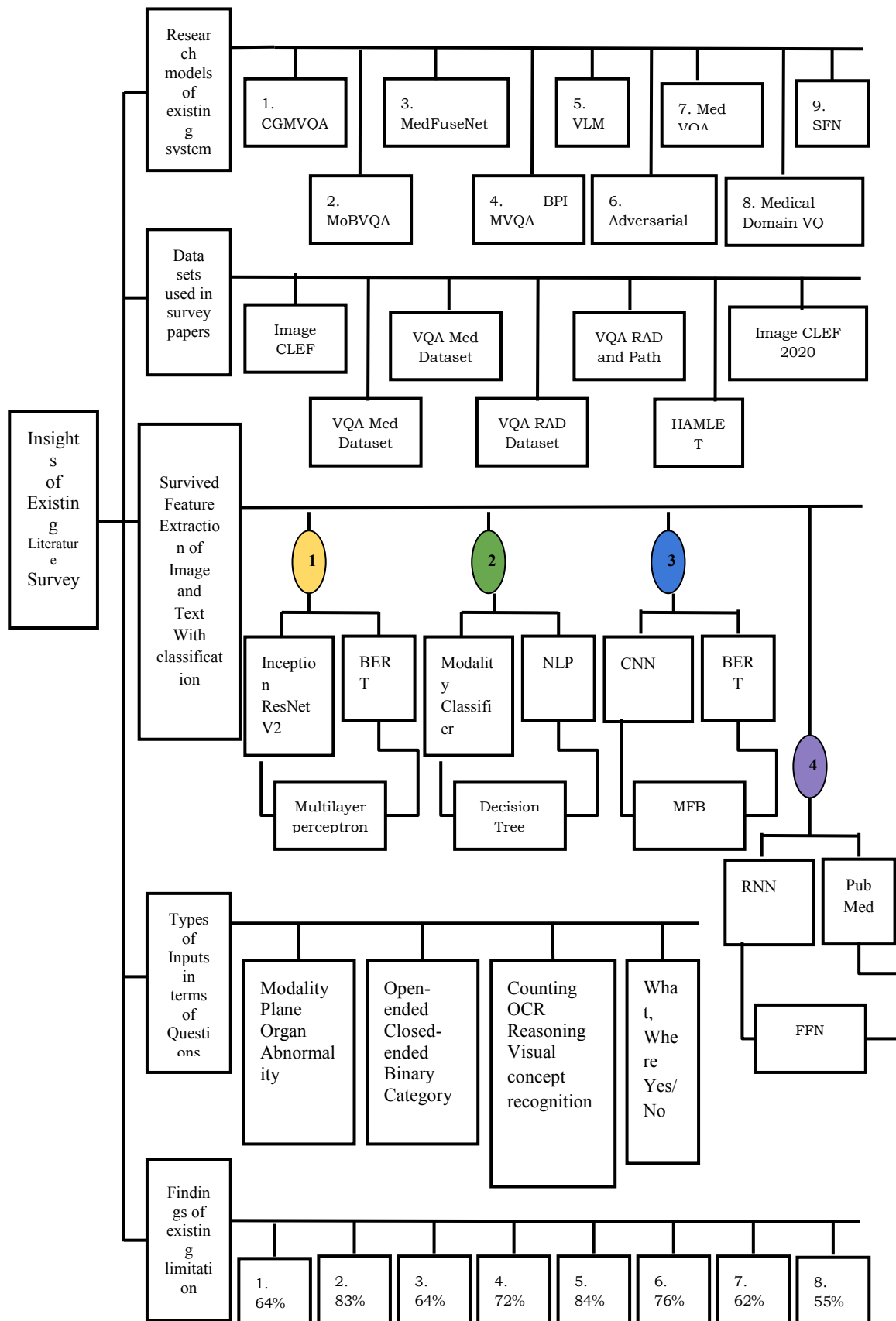


Fig 5 : Insight of Literature Survey

2.2 Identification of challenges and gaps in existing research

Preparing numerous appropriate queries for the image while integrating both visual and textual information is a challenging task. These questions should be categorized as near end, open end, descriptive, summary, etc. High resolution is essential when examining radiology images, and [31] enumerates various types of radiological images, each differing in shape, size, quality, and depiction. For each image, a natural language question is provided, and the objective is to analyze image attributes and delineate a bounding box around the object that answers the question. Several challenges arise in this context:

Object Detection Limitations: Object detection faces two types of limitations, namely object boundary boxes and non-object boundary boxes. Object recognition restricts the subset of the object, localizing and labeling the boundary box for all relevant images.

Semantic Segmentation Challenges: Finding and recognizing objects in an image or video sequence, performing semantic segmentation on meaningful images, and classifying them into a specified class while labeling each pixel with the class item pose as some of the more challenging tasks in VQA datasets.

Effectively assessing the quality of brief answers, as well as answers with significant variation in spelling, phrasing, and grammar, poses a challenge.

2.2.1 Challenges in VQA System

Training accurate and reliable VQA models requires large datasets with diverse and well-annotated medical images and corresponding questions. Developing extensive datasets in the healthcare domain is difficult due to privacy concerns, the necessity for expert annotations, and the wide range of medical disorders. Medical questions often involve complex scenarios, requiring a deep understanding of anatomical structures, clinical context, and nuanced information. VQA models may struggle with the complexity of medical queries, especially when dealing with rare conditions or questions that demand a profound medical knowledge base. VQA models trained on specific datasets may struggle to generalize well to new, unseen scenarios or diverse medical imaging practices. The lack of standardization in medical imaging and the diversity of clinical settings can hinder the generalizability of VQA models. VQA

models heavily rely on the quality of features extracted from medical images. Inaccuracies in feature extraction, especially in complex images like MRIs or CT scans, can lead to incorrect answers. Improving feature extraction methods is crucial. Implementing Visual QA in healthcare entails resolving ethical problems such as patient privacy, data security, and model biases. Adhering to strict ethical standards and compliance with healthcare regulations becomes paramount, adding complexity to the deployment of VQA systems. Embedding VQA systems seamlessly into clinical workflows poses a significant challenge. Clinicians often work with a variety of tools, and integrating VQA into existing systems without disrupting clinical processes requires careful consideration. Interpreting and explaining the decisions of VQA models is crucial in healthcare for gaining the trust of medical professionals. Black-box models may be met with skepticism in healthcare, where understanding the reasoning behind a decision is crucial for acceptance. VQA systems may require significant computational resources, impacting real-time processing. In healthcare, especially during critical decision-making processes, delays caused by resource-intensive VQA models could be detrimental. The medical field is dynamic, with ongoing discoveries and updates to medical knowledge. VQA models may become outdated if not regularly updated to incorporate the latest medical findings and practices. Medical questions often involve uncertainty, requiring models to provide nuanced and probabilistic responses. VQA models need to incorporate mechanisms to handle uncertainty and express confidence levels in their answers. Understanding and mitigating these limitations are critical for the responsible and effective deployment of Visual Question Answering systems in the healthcare sector.

2.2.2 Inhibitions of Datasets

Medical datasets for VQA are often smaller and less diverse compared to general VQA datasets. This limitation arises due to challenges in collecting and annotating medical images. Medical images are sensitive and subject to strict privacy regulations. Obtaining consent and anonymizing data while maintaining its usefulness for training can be challenging. Annotating medical images and generating meaningful questions can be more complex than in other domains. Domain expertise is required, making the annotation process labor-intensive and potentially prone to errors. Medical imaging encompasses various modalities such as X-rays, MRIs, CT scans, etc. Building a

comprehensive dataset that covers these modalities requires significant effort and collaboration with healthcare institutions. Rare medical conditions may not have sufficient annotated examples in datasets, limiting the model's ability to handle queries related to uncommon diseases. Different medical professionals may interpret images differently, leading to inter-observer variability. This variability can introduce ambiguity in annotations, affecting the reliability of the dataset. Medical conditions often involve changes over time. Capturing longitudinal data and ensuring that datasets represent the temporal evolution of diseases is a challenge. Lack of standardization in medical imaging practices and formats can hinder the creation of uniform datasets.

Harmonizing data across institutions is difficult due to variations in equipment and protocols. The data collection process may inadvertently introduce biases, such as over-representation of certain demographics or conditions, impacting the generalizability of the model. Detailed annotations at a fine-grained level, such as lesion boundaries or specific anatomical structures, are often lacking. This limits the potential for training models for specific diagnostic tasks. Domain shifts may cause models trained on one dataset to not generalize adequately to new datasets, particularly if the datasets come from different healthcare institutions with variations in imaging protocols. VQA datasets may not adequately represent the complexity of real-world clinical scenarios, which involve interacting with patients, understanding diverse medical histories, and considering various contextual factors. Overcoming these inhibitions requires collaborative efforts between the AI research community, healthcare professionals, and institutions to create diverse, representative, and ethically sourced datasets for training robust VQA models in the medical domain.

2.2.3 Drawbacks on various techniques

Clinical requirements for developing practical and effective applications present six crucial challenges: Question heterogeneity, additional healthcare information, comprehension, extrapolation, utilization of high language models, and seamless fusion into the healthcare workflow. These challenges are proposed to inspire researchers to develop mature and accurate medical Visual or image QA systems that can significantly contribute to clinical decision-making [4].

2.3 Gap Identification from Existing Research

Several innovative methods have emerged for addressing Visual Question Answering (VQA) tasks, driven by the intriguing challenges presented in Healthcare Image QA. These strategies hold potential for application in the medical domain, although Healthcare Image QA is still in its initial stages of development. Prior to Healthcare Image QA, the medical domain already had question-and-answer (QA) systems primarily employed for databases, information retrieval, and other technologies.

The planned research initiative aims to develop an advanced Visual or image Question Answering (QA) system with the potential to offer significant societal benefits. The primary goal is to develop an optimal methodology for extracting image and text features from radiological images, with a focus on high accuracy and outperforming existing methods. To achieve this goal, the selective search technique is employed to create about 2,000 region proposals from the input images, which are then downsized to a predetermined, set size. The following initiatives collect a feature vector of length 4,096 from each region suggestion. Finally, a pre-trained Support Vector Machine (SVM) algorithm is employed in the third module to classify each region proposal into either the background or any of the object classifications.

The Kaiming initialization method is used in the research project to extract textual features, allowing extensively layered models (over 30 layers) to converge effectively by precisely modeling the ReLU non-linearity. The ideal weight distribution following ReLU would have a little higher mean layer by layer and a variance close to one. To do this, the weight initialization uses a normal distribution with a mean of zero and a variance of one.

By combining these carefully designed strategies and techniques, the research project aims to build an advanced VQA system for medical images, contributing to improved medical decision-making and diagnosis.

In a technologically advanced society, operating an automated Visual or imaginary QA system in the health domain is a tedious task, as users require accurate responses to questions about medical images. Since it involves people's health, ensuring precise communication becomes crucial. Therefore, this research aims to propose an automated system that can accurately answer user queries related to medical images.

One of the significant challenges in this context is the presence of numerous complex medical terms that users may find difficult to comprehend. To address this issue, the research will focus on creating a dictionary of medical terms to aid the VQA system. This dictionary will play a significant role in enriching the system's ability to provide relevant and easily understandable answers.

The study will introduce a groundbreaking algorithm for dictionary creation that will encompass both image and Question-Answer (QA) pairs. By combining the information from these pairs, the algorithm will effectively compile a comprehensive and contextually relevant medical dictionary.

To assess the proposed system's performance, numerous models will be compared using various datasets. The goal is to select the model that performs best and improves the overall system.

The research will also analyze the proposed methodology's performance in comparison to existing techniques. This investigation attempts to assess the proposed approach's uniqueness and efficacy in providing accurate and reliable replies to medical image-related inquiries.

Identifying the gaps in a Visual QA (VQA) system specifically designed for healthcare images involves recognizing areas where the current system falls short or lacks adequate solutions. Here are some key gaps that can be addressed to improve the VQA system for medical images:

1. Limited Medical Domain Expertise:
 - Many existing VQA systems for medical images are developed by computer vision and natural language processing experts without extensive healthcare domain knowledge.
 - There is a need to collaborate with medical professionals to ensure accurate understanding of medical images and proper annotation of questions and answers.
2. Lack of Large-Scale Medical VQA Datasets:
 - Creating large-scale datasets with diverse medical images, relevant questions, and accurate answers is challenging due to privacy and ethical concerns.

- Efforts should be made to curate comprehensive and representative medical VQA datasets for training and evaluation.
- 3. Addressing Biases in Medical VQA:
 - Biases in the data can lead to biased predictions in VQA systems, affecting fairness and trustworthiness.
 - Identifying and mitigating biases in the medical VQA system is essential to ensuring equitable and reliable performance.
- 4. Handling Ambiguity in Medical Questions:
 - Medical questions can be complex and ambiguous, Necessitating profound comprehension of medical context and expertise in domain-specific knowledge.
 - The VQA system needs to handle such ambiguity effectively to provide accurate and informative answers.
- 5. Explainability and Interpretability:
 - Medical professionals often need explanations for the system's predictions to trust and validate the results.
 - Developing explainable VQA models that provide clear reasoning for their answers is critical in medical settings.
- 6. Integration with Electronic Health Records (EHRs):
 - Integrating the VQA system with EHRs could enhance clinical decision-making and streamline the diagnostic process.
 - However, challenges related to privacy, data sharing, and compatibility with different EHR systems need to be addressed.
- 7. Handling Rare or Unseen Medical Conditions:
 - Medical images may contain rare or previously unseen conditions that the VQA system might struggle to recognize.
 - Strategies like transfer learning or meta-learning could be explored to improve performance in rare cases.
- 8. Multimodal Fusion for Medical VQA:
 - Efficiently fusing information from medical images and textual questions remains a challenge in VQA systems.
 - Investigating advanced multimodal fusion techniques to capture the relationship between medical images and textual context is essential.

9. Adapting to Multilingual Settings:

- In a diverse medical environment, the VQA system should be adaptable to answer questions in multiple languages.
- Extending the system to handle multilingual settings can improve accessibility and usability.

Addressing these gaps will lead to more robust and accurate VQA systems for medical images, providing valuable support to medical professionals in clinical decision-making and advancing the field of medical image analysis.

METHODOLOGY



In the domain of visual question answering (VQA), the computational task involves presenting a computer with an image along with a relevant question, eliciting a response that adequately addresses the inquiry. A longstanding objective in the sphere of artificial intelligence research has been the development of machines capable of comprehending visual content and providing responses akin to human understanding. Notably, visual question answering (VQA) has emerged as a prominent academic discipline. Within the specific domain of medical visual question answering (Med-VQA), clinical queries are accompanied by radiological images, aiming to devise a system that can accurately generate responses based on the visual data contained in the images.

3.1 Description of the research design and methodology used

The system architecture for a Visual QA (VQA) system in the context of skeletal images typically involves several key components designed to process and understand both skeletal image and textual knowledge in the current era. Here's an overview of a generalized architecture with Input module their Image and question as the input, This module handles the medical images. It may utilize various techniques like Convolutional NN (CNNs) for visual feature extraction. Pre-trained models like ResNet or VGG might be employed for this purpose. The textual input includes questions related to the medical images. Natural Language Processing (NLP) approaches are often used to preprocess and encode the textual data. Image and text features are extracted separately. For images, CNNs are employed to capture visual patterns and features. For text, embedding layers and recurrent networks like Long Short-Term Memory (LSTM) or bidirectional LSTMs may be used to understand the context and relationships in the question.

Multimodal fusion pertains to the amalgamation of extracted features from both modalities, namely images and text. This fusion can occur at various levels, encompassing concatenative fusion, which requires integrating features at a basic level, and subsequent fusion, which integrates features at a higher, more abstract level. Techniques for multimodal fusion include concatenation, element-wise multiplication, and attention processes. The resultant fused features may undergo further processing, potentially within additional neural network layers. This stage facilitates the comprehension of intricate correlations between visual and textual

information. Ultimately, the final layer is tasked with making predictions or guesses, with a softmax layer commonly employed in classification challenges. In regression tasks, a linear layer is commonly employed. The training of the entire system involves utilizing a dataset consisting of paired medical images, associated questions, and corresponding responses. During training, the network's parameters are optimized to minimize the disparity between expected and actual answers. Following training, the system undergoes evaluation on a separate dataset to assess its effectiveness, frequently assessed using measures like accuracy, precision, recall, and F1 score.

Post-processing techniques may be implemented based on specific needs, which could involve refining the answer or providing additional context.

This design is adaptable and can be tailored to unique needs and the nature of the medical VQA assignment at hand. Various model topologies, pre-processing processes, and fusion procedures can be investigated to optimize performance.

This section underscores the diverse visual feature extraction methods available for distinct datasets, with a specific emphasis on medical radiological images. The primary objective of the research is to proficiently train models using a variety of medical images, queries, and associated responses. To accomplish this, two distinct feature extraction methodologies were employed for the dataset.

The feature extraction process involves two pivotal components—visual and textual. In the context of this proposal, the focus is directed towards medical radiological images. The datasets are divided into a 70:30 training-to-testing ratio, which means that 70% of the dataset is allocated for training, while the remaining 30% is allocated for testing.

There are various models of processing mentioned in this chapter, which are highlighted below.

User Interaction: The user initiates the process by inputting a medical image into the model along with relevant questions.

Image Feature Extraction: The system meticulously extracts features from the input images, diligently searching for distinctive patterns.

Text Feature Extraction: Simultaneously, the text entered as a question undergoes feature extraction, ensuring that relevant textual features are identified.

Question Classification and Prediction: The loaded question is subjected to classification, and the model employs predictive analysis to furnish accurate responses.

This formal description elucidates the systematic approach adopted in the research, encompassing the selection of feature extraction methods, dataset distribution, and the intricate processes involved in model training and prediction.

3.1.1 Morphology of a Radiology Image

Radiology is a specialized field within the medical profession that employs various imaging techniques to detect, diagnose, and address a spectrum of disorders [63]. Within radiology, there are two key subspecialties: diagnostic radiology and interventional radiology [64].

Diagnostic Radiology:

Diagnostic radiology enables radiologists to meticulously examine internal bodily structures, facilitating the identification of the root causes of symptoms, screening for health issues, and monitoring the body's response to treatment. Common modalities in diagnostic radiology encompass plain radiographic images, computed axial tomography (CT), magnetic resonance tomography, positron emission tomography (PET), and ultrasound imaging [65]. These modalities are adept at visualizing a diverse range of ailments, including but not limited to breast cancer, colon cancer, and heart disease.

Diagnostic Imaging Modalities:

CT (Computerised Tomography): Often referred to as CAT (Computerised Axial Tomography), CT is a widely utilized diagnostic radiology examination. It includes various applications such as CT angiography, fluoroscopy with upper gastrointestinal (GI) studies, MRI (Magnetic Resonance Imaging) and MRA (Magnetic Resonance Angiography) scans, mammography, bone scans, thyroid scans, plain x-rays, PET (Positron Emission Tomography) images, PET scans, PET-CT scans, and ultrasound [64, 31].

System Functionality:

When a user submits a medical image to the system, the system conducts a comprehensive comparison with its database. Subsequently, the system provides the user with pertinent information crucial for guiding the next phase of medical analysis and decision-making. This formal elucidation underscores the significance of diagnostic radiology in medical practice, elucidating its various modalities and their essential roles in disease detection and characterization.

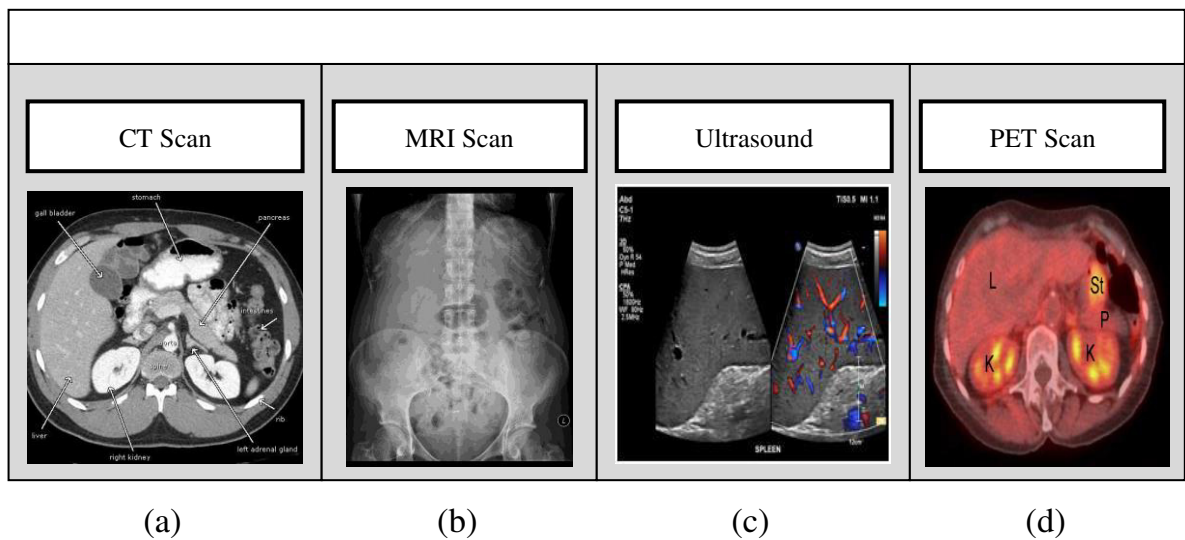


Fig 6 : Types of Radiology Image

The above figure indicates the different types of radiology images used for VQA systems. Each image describes a different description in a different angle. Based on the image the modality of question and answer will change. Fig (b) is from [66] which has lots of information about it, same for other images too. So collect a huge amount of information regarding images, question and answer. Remaining chapters elaborate the significance and methodology of current vqa system, architecture, uses, its error rate, accuracy. Then finally our proposed system projects the limitations of other models and introduces the new model to answer the question with high ratings.

Specifically, many Visual or imaginary QA models have focused on how they incorporate question and skeletal inputs into the model. Various Visual or image QA techniques were reviewed in the previous chapters. Different datasets used for various methodology, based on the problem state, the models will differ.

3.2 Explanation of data collection methods, tools, and procedures

As of the last knowledge update in September 2022, several Visual or imaginary QA (VQA) models have been proposed and applied in the healthcare sector. However, the field is dynamic, and new models may have been developed since then. Here are some notable VQA models that were relevant in healthcare. MedVQA is specifically designed for radiology visual question answering. The system employs convolutional neural networks (CNNs) to analyze images and recurrent neural networks (RNNs) to handle textual queries. Inspired by BERT (Bidirectional Encoder Representations from Transformers), Medical-BERT utilizes transformer-based models to respond to queries regarding both textual and visual content in medical images. The MQR-VQA (Multi-Modal Quality-Aware Relevance) model aims to assess the relevance and quality of regions in medical images to answer questions. It combines visual attention mechanisms with textual information. The MedVQA-Transformer model employs transformer architectures for processing both visual and textual information. Recognized for their effectiveness in capturing long-range dependencies in sequences. DeepMedQA uses a combination of deep learning techniques, including CNNs for image processing and recurrent networks for text. It is designed to answer questions related to medical images. The Attention-Gated CNN-LSTM model integrates attention mechanisms into a network composed of convolutional neural networks (CNNs) and long short-term memory (LSTM) units. The attention mechanism directs the model's attention to relevant sections of medical imagery. Healthcare Image QA is a framework developed for skeletal visual or image question answering systems. It utilizes deep learning models to process both visual and textual data, enabling the model to answer questions about medical images. Remember to check for recent publications, conferences, or preprint archives for the latest advancements in the field. Additionally, the specific requirements of a healthcare application may lead to the development of specialized models tailored to particular tasks or medical domains.

In this section, we are looking forward to current methodologies such as Linear classifier, k-nearest neighbors algorithm, Softmax classifier, support vector machine, Convolutional Neural Network (CNN), regions with convolutional neural networks (R-CNN), fast regions with convolutional neural networks (FR-CNN), Faster regions

with convolutional neural networks (FR-CNN), Deep Belief Networks (DBN), Long short-term memory (LSTM) AND Bidirectional Long short-term memory (BiLSTM).

3.3 Current Methodology on Both Visual and Textual Feature Extraction Techniques

3.3.1 Linear Classifier

The purpose of this document is to investigate and analyze the application of linear classifiers in the domain of skeletal imaginary questions and answers specifically tailored to medical images. The focus will be on understanding how linear classifiers, known for their simplicity and interpretability, can contribute to answering questions related to complex medical visuals. Through a thorough exploration, we aim to uncover the strengths, limitations, and potential advancements that linear classifiers bring to the challenging task of VQA within the medical context. This exploration encompasses the entire pipeline, from the extraction of features from healthcare related images to the design, training, and optimization of linear classifiers, with a keen eye on their practical applications and performance evaluation metrics in the medical domain. Ultimately, this document seeks to provide insights into the role of linear classifiers as a valuable tool in enhancing the interpretability and accuracy of VQA systems when dealing with medical imagery.

Linear classifiers are a category of machine learning models designed to categorize input data points into discrete groups or classes. The fundamental concept behind linear classifiers is based on the idea of drawing a decision boundary in the feature space, which separates the data instances belonging to various classes. This decision boundary is a hyperplane, a subspace with one dimension less than the input space, and is determined by a set of parameters.

The linear classifier's decision-making process involves categorizing an input based on which side of the decision boundary the point falls on. The decision boundary is defined by a linear combination of the input features, each multiplied by a weight, plus an additional bias term. Mathematically, it can be represented as:

$$f(x) = \text{sign}\left(\sum_{i=1}^n w_i \cdot x_i + b\right) \quad (1)$$

Here:

- $f(x)$ is the decision function.
- x_i represents the input features.
- w_i represents the weight associated with each feature.
- B represents the intercept or bias term
- Sign is the sign function, determining the class label based on the sign of the expression.

Training a linear classifier entails determining the best values for the weights and bias. This is typically achieved through optimization algorithms that aim to minimize a certain objective function, often associated with the misclassification of training data. Key characteristics of linear classifiers include simplicity, interpretability, and efficiency. They work well when the relationship between features and classes is approximately linear. However, they may struggle with more complex, non-linear relationships in the data. Linear classifiers are frequently employed across diverse applications, including picture classification, text categorization, and, depending on your environment, visual or picture question answering (VQA). In the healthcare domain, linear classifiers can be useful for tasks such as disease diagnosis using visual data.

3.3.2 Traditional Neural Network

Suppose if the image pixel size is 1000×1000 , then there will be a total of 3 (rgb size) $\times 1000 \times 1000$ features, which means 3 million input features are there. The first layer of the neural network will have 1000 input characteristics that reflect the neural network's weight, or a connected layer of two layers. So the whole amount of weight is given below.

No. of weights = $3 \times 10^6 \times 10^3$ which means 3×10^9

So, there are 3 billion weight parameters in a single layer.

The size is too large to load into the system and model. Also, it is very difficult to manage using laptops or our PCs because of this drawback. The time requirement to train the model is very high. The traditional neural network is overfitting for the

object detection concept. To overcome this drawback, a new technique arrived on the market that we called the Convolutional Neural Network (CNN).

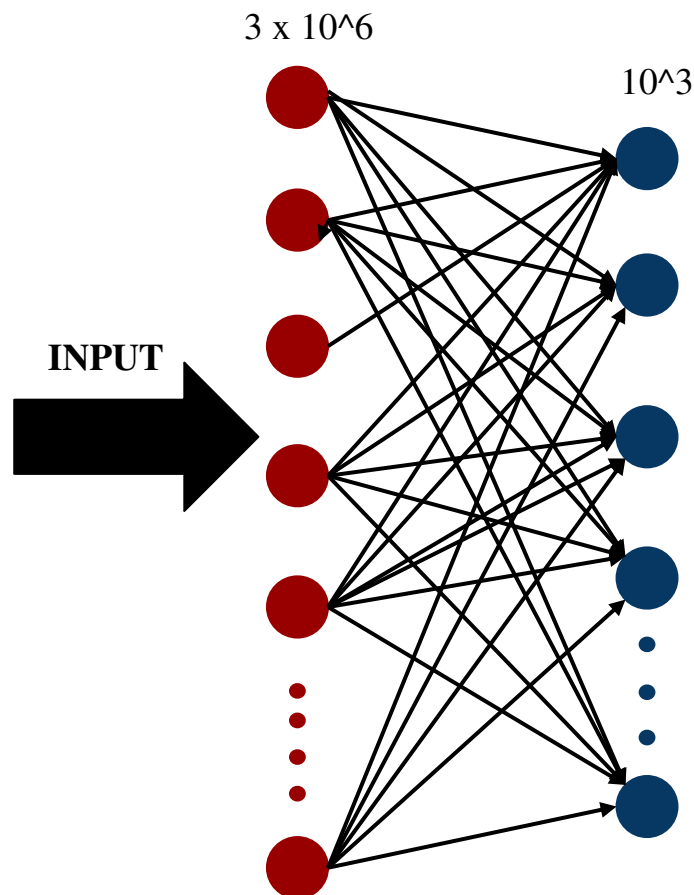


Fig 7 : Traditional Neural Network with a Single Layer

3.3.3 The prime Idea behind a Convolutional Neural Network for image feature extraction

The primary idea behind a neural network that uses convolution is to use filters. These filters are located on sliding windows. This filter is tasked with identifying the characteristics of objects and patterns in the image



Fig: 8 : RGB Image and Gray scale Image to find features of the image

The features of the above image are Shape, Size, Edges, Colors etc. We can use the images with vertical edge detector and horizontal edge detector, when combine then it will be 3×3 pixel filter which means 9 pixel size, Thus we reduce the number of parameters from the objects.

1	0	-1
1	0	-1
1	0	-1

Fig 9 : 3×3 pixel object

Features of Human face

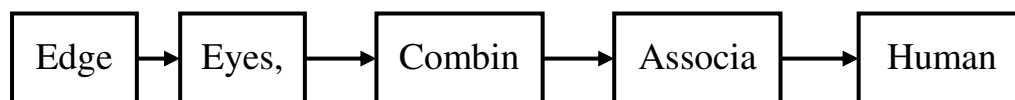


Fig 10 : Structure of CNN process from Input object to final output object with feature extraction

The RGB image converts to a black and white image; every pixel value starts at 0, which means black color, and ends with 255, which means white.

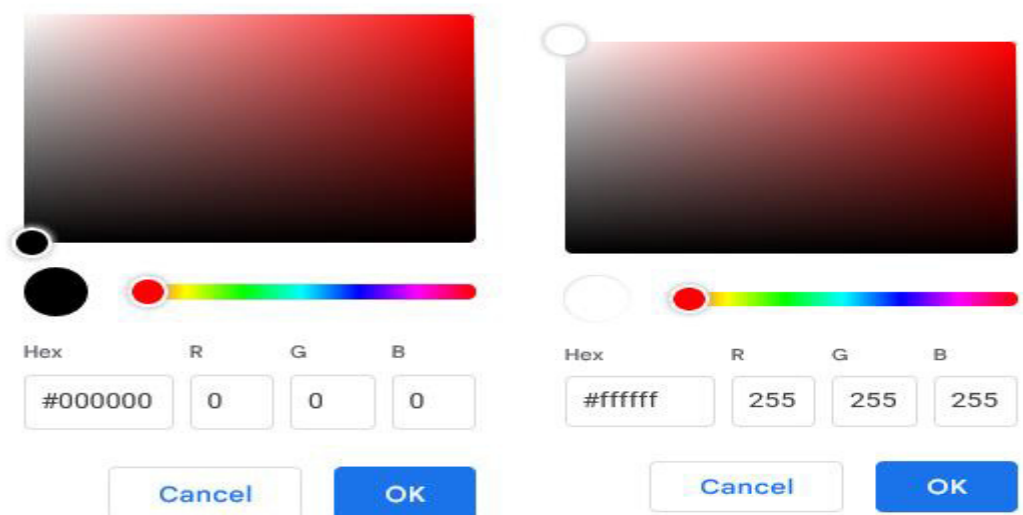


Fig 11 : The value of Black and white colors 0 to 255

Example:

6×6 Pixel images convolution operation with 3×3 (RGB), then the Output Image size will be 4×4 Pixel.

1	6	9	10	2	8
2	5	1	8	4	2
3	7	4	9	10	3
9	8	3	6	7	9
8	0	9	4	7	2
9	10	12	6	9	8

 $*$

1	0	-1
1	0	-1
1	0	-1

 $=$

-8	-9	-2	14
6	-3	-13	9
4	-4	-8	5
2	2	1	-3

Fig 12 : How the convolution operation take place with 6×6 Pixel images and 3×3 image

The output result calculated as for first pixel is $(1 \times 1) + (2 \times 1) + (3 \times 1) + (6 \times 0) + (5 \times 0) + (7 \times 0) + (9 \times (-1)) + (1 \times (-1)) + (4 \times (-1)) = -8$

$$(n \times n) * (f \times f) = (n - f + 1) \times (n - f + 1).$$


Fig 13 : Black and white input image for convolution operation to identify the vertical edge

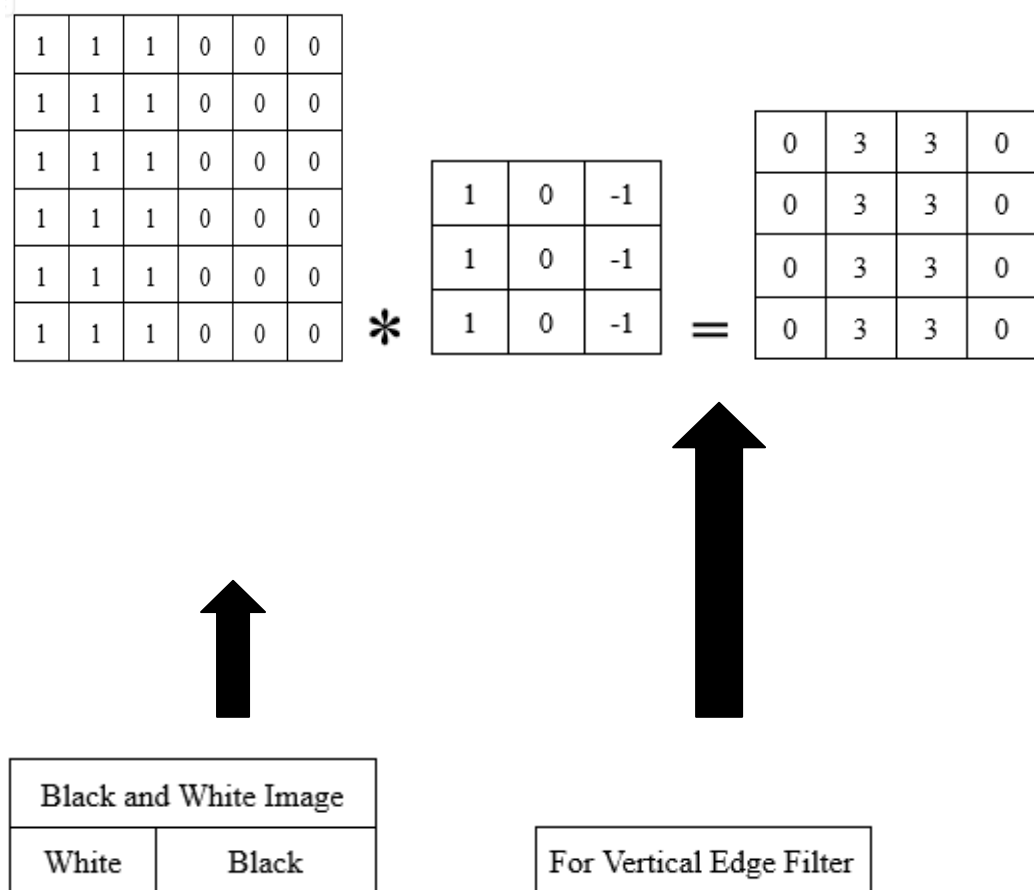


Fig 14 : Black and White image with 3 × 3 px vertical edge filter the output is in

4 × 4px

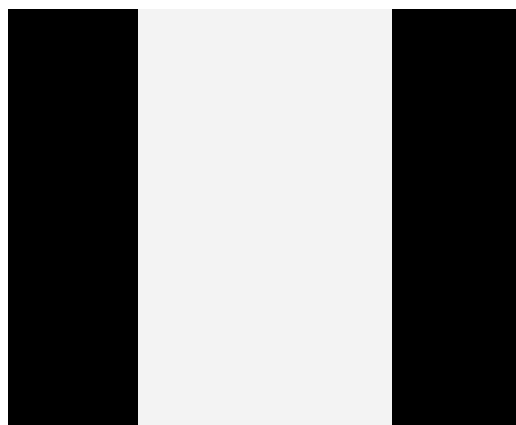


Fig 15 : Output image from above figure after 3 × 3 px convolution with 6 × 6 px

The output result calculated as for first pixel is $(1 \times 1) + (1 \times 1) + (1 \times 1) + (1 \times 0) + (1 \times 0) + (1 \times (-1)) + (1 \times (-1)) + (1 \times (-1)) = 0$

The remaining calculation will be done the same way for other pixels.

1	1	1
0	0	0
-1	-1	-1

Fig 16 : Image 3×3 px horizontal edge filter

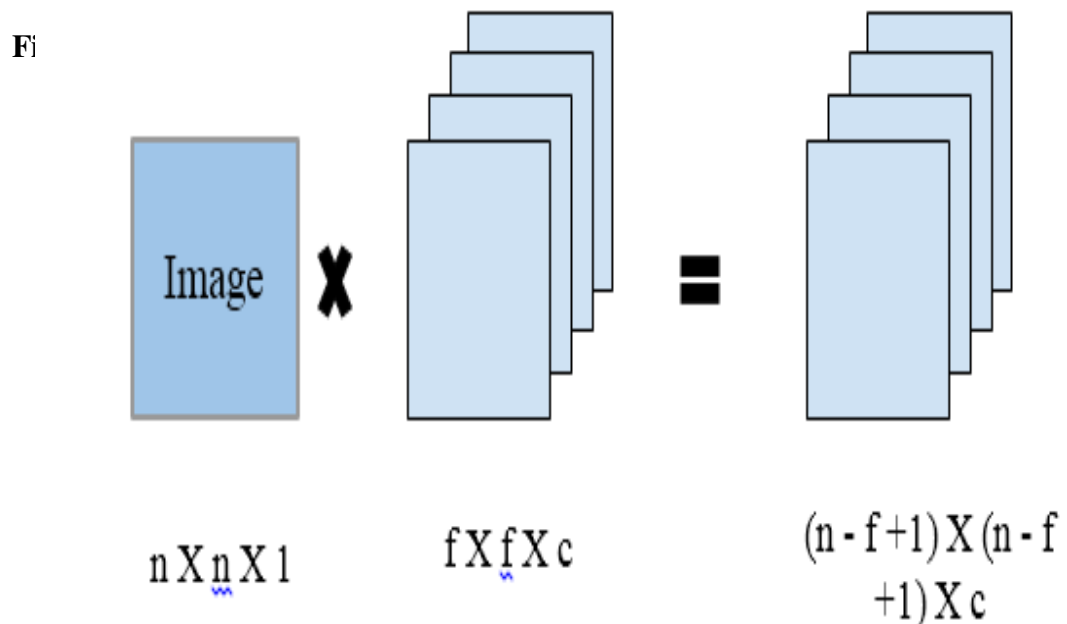


Fig: 17. $n \times n \times 1$ for Gray scale image with $f \times f \times c$ number of filters the output will be multiple images.

When dealing with a colored or RGB image with dimensions $n \times n \times 3$, the convolution process involves using a three-channel filter with dimensions $f \times f \times 3$. Each channel of the filter aligns with the corresponding channel of the input image. The filter, being a 3D cube, consists of a total of 27 values (3 values per channel, totaling 9 values per layer). To perform convolution between the input image and the filter, we overlay the two images and multiply the values in each corresponding cell. These products are then added together to form a single output value, which is displayed in the first cell of the output image. This operation is repeated, shifting the filter one pixel to the right until the entire input image is covered.

When converging an $n \times n \times 3$ image with a $f \times f \times 3$ filter, the resulting output image will have dimensions $(n-f+1) \times (n-f+1)$. This output represents a single image resulting from the convolution operation. In a convolutional neural network (CNN), multiple such filters are utilized in a single layer. Each filter may have different parameter values, which are acquired throughout the training procedure. For instance, in the example provided, a filter with hardcoded values such as (1, 1, 1), (0, 0, 0), and (-1, -1, -1) may behave as a vertical filter.

However, during training, these parameter values will adjust autonomously, allowing the filters to recognize relevant features in the images. The process of parameter adjustment during training enables the filters to effectively identify and extract meaningful patterns from the input data.

In convolutional neural networks (CNNs), the parameters of the filters are adjusted during the training process, enabling them to adapt and identify specific features within input images. This capacity allows CNNs to accomplish tasks like image categorization, object identification, and image segmentation.

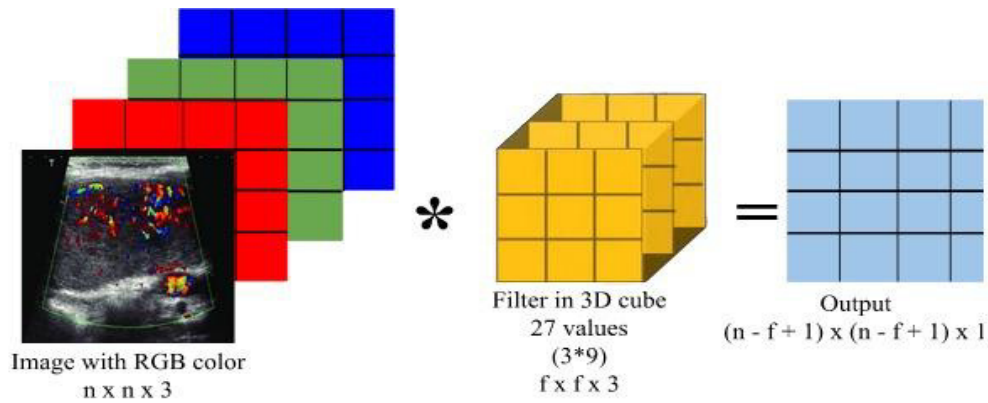


Fig: 18. Single filter convolution with single output

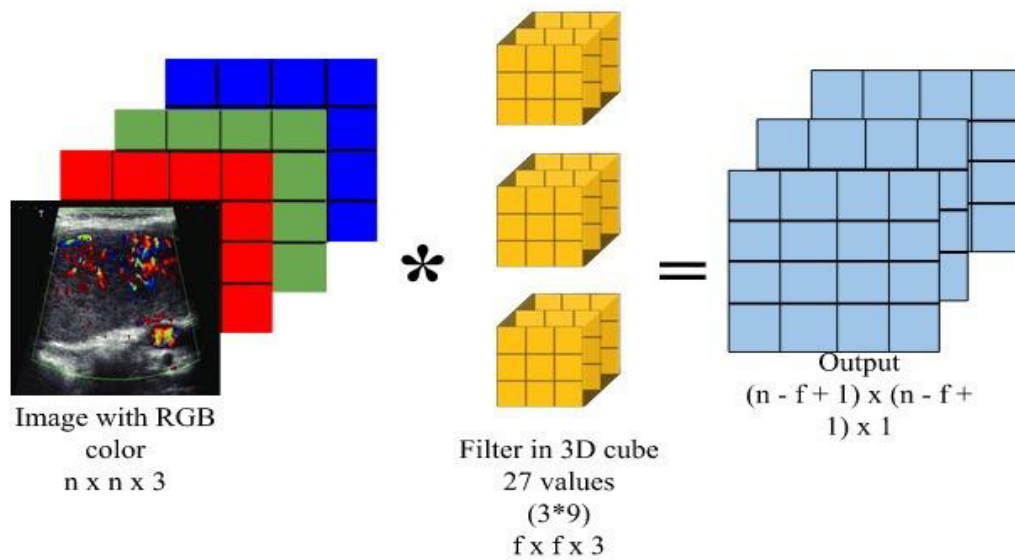


Fig: 19. Multiple filter convolution with multiple output

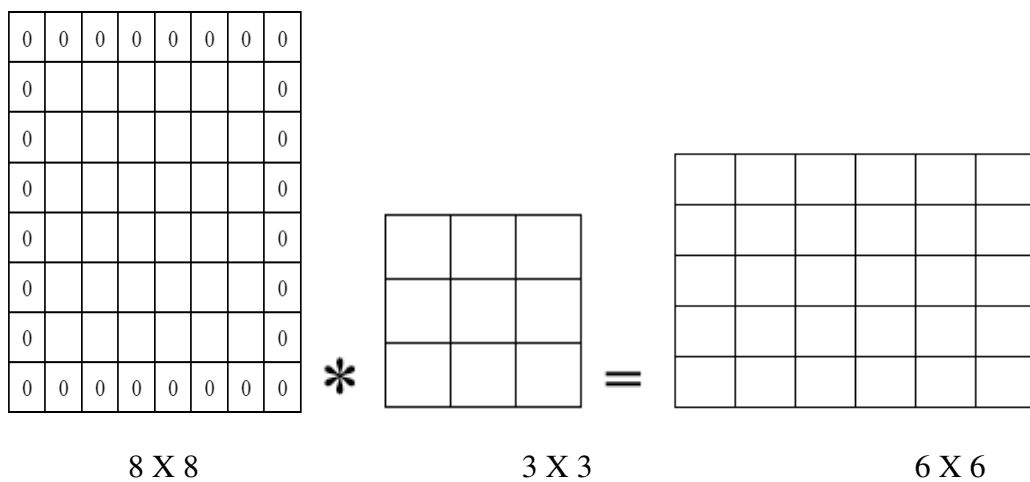


Fig 20 : Padding in Convolutional Neural Network

There are two main choices when it comes to padding VALID and SAME. When VALID performs the convolution operation with no padding at all,. After performing the convolution operation, the output image should be the same size as the input image. The size of new image will be

$n' = n + 2p$, where p is amount of padding that we are applying

$(n' - f + 1) = n$ that is $(n + 2p - f + 1) = n$ the value of p is given below

$$p = \left(\frac{f - 1}{2}\right)$$

If we use the Tensor Flow framework and apply a convolution layer, we just specify what kind of padding we want.

`tf.nn.conv2D(X, W1, strides = [1, 1, 1, 1], padding = 'VALID')`. Here 'f' is usually taken as an odd number. If we take an even number, then we will need to apply uneven or asymmetric padding.

For example, we might need 1 pixel of padding on one side of the image, and on the other side, we might need 2 pixels.

Max Pooling is the most powerful operation on CNN. There can be many types of pooling layers. In CNN, the function pooling layer is to reduce the size of the dimension of the image by preserving the feature on it. For the max pooling operation, we need to fix a filter for a particular size and stride a particular amount. This filter can be of any size or amount. Usually, any amount of stride is taken from the same size as the side length of the filter. For example, the filter size is 2×2 and the stride is

2. After the performance, the output will be an image with a 2×2 matrix. Performing the max pooling operation, we will see the filter image in the top left column of the input image, then we will extract the maximum value of the window where the filter image is as per stride size for the next maximum value. This will be repeated till the last pixel. Why might we want to do this max pool?

The first reason is that it reduces the image size and, thus, the computational cost. So that we can get the output faster and train the model faster. We want to do our operations quickly. Even though it reduces the size of the image, it still preserves the features of that image, and max pooling works in such a way that it sharpens the

features or enhances the features. Suppose the input image has a vertical edge; this vertical edge is still preserved but actually enhanced in the output image. As shown in one window of an image, we can say that the maximum value holds the maximum intensity of the features. When they extract the maximum value, it sharpens the features. Now where do we need max pooling in CNN?

The standard procedure in convolutional neural networks (CNNs) involves applying a max pooling layer after the convolutional layer. This step effectively reduces the size of the output image from the convolutional layer while also enhancing the extracted features. The convolution layer itself executes the convolution process using multiple filters. Suppose if we use 5 numbers of filters in one convolution layer then we know that it will generate 5 numbers of output images. Thus after performing the max pooling operation we will again get the same number of output images. Which means 5 numbers of images in the output of the max pooling layer. From the entire CNN we will use many convolution layers and max pooling layers. Here there is a small side note that it does not always use max pooling layer after every convolution layer. We may use convolution layers and may skip max pooling layers or just fewer numbers of max pooling layers than the number of convolution layers. The reason for that is the max pooling layer reduces the size of the image but we might not reduce the size too much, if we are using a very big convolutional neural network. The use of max pooling will improve the performance of CNN. There are no parameters in it or no training in it. Another pooling method is average pooling, which takes the average of values filter size and image size pixel.

To perform the convolutional layer we use the convolutional operation of the input radiology image to get the value of the input skeletal image which leads to it being scaled properly. So, we add a 'bias' which denotes 'b' and assign it to the nonlinear function. This non linear function can be tanh or ReLU but most commonly ReLU is used after that we will get the output as size as compared with input image. So the entire convolution is one applying the convolution operation and one applying ReLU operation or nonlinear activation function. These combain we called convolution layers.

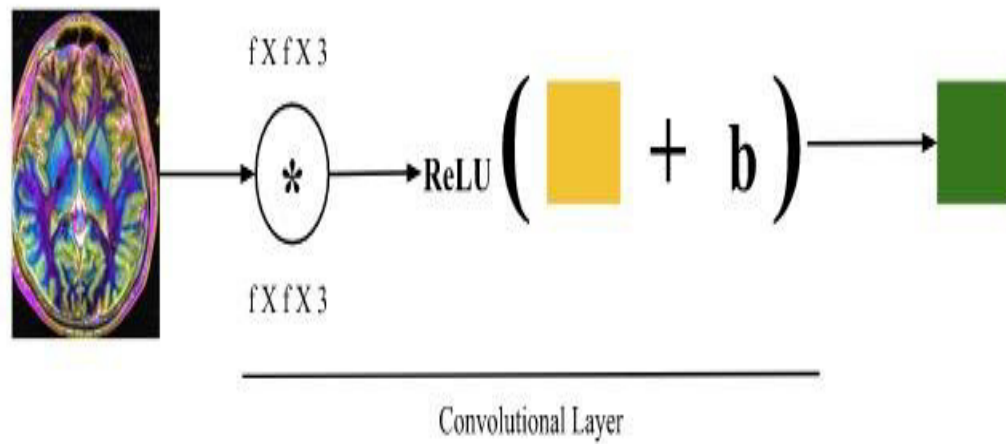


Fig: 21. Single Filter Convolution layer using non linear function and bias with single image as output

If we use multiple layers, the bias will also be changed, and we will get multiple outputs, as shown in the below figure.

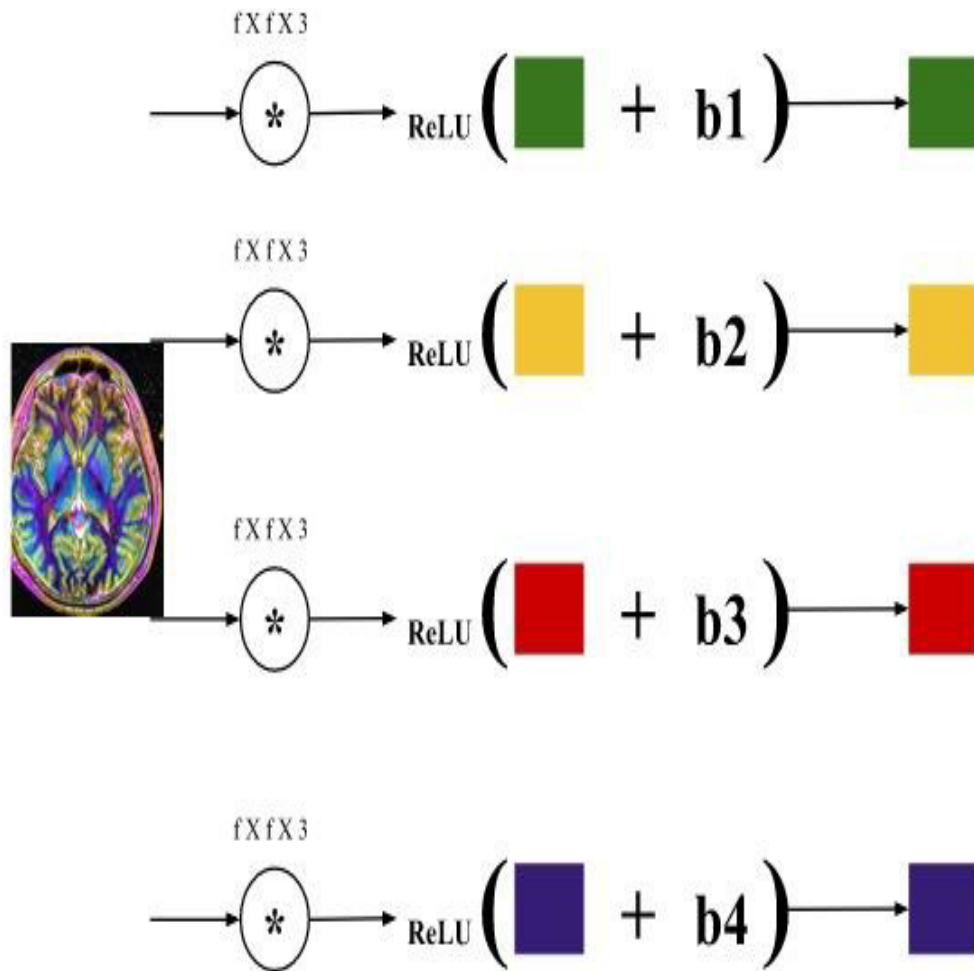


Fig. 22. Multiple Filter Convolution layer using non linear function and bias with multiple output

So, in short, we convert it as,

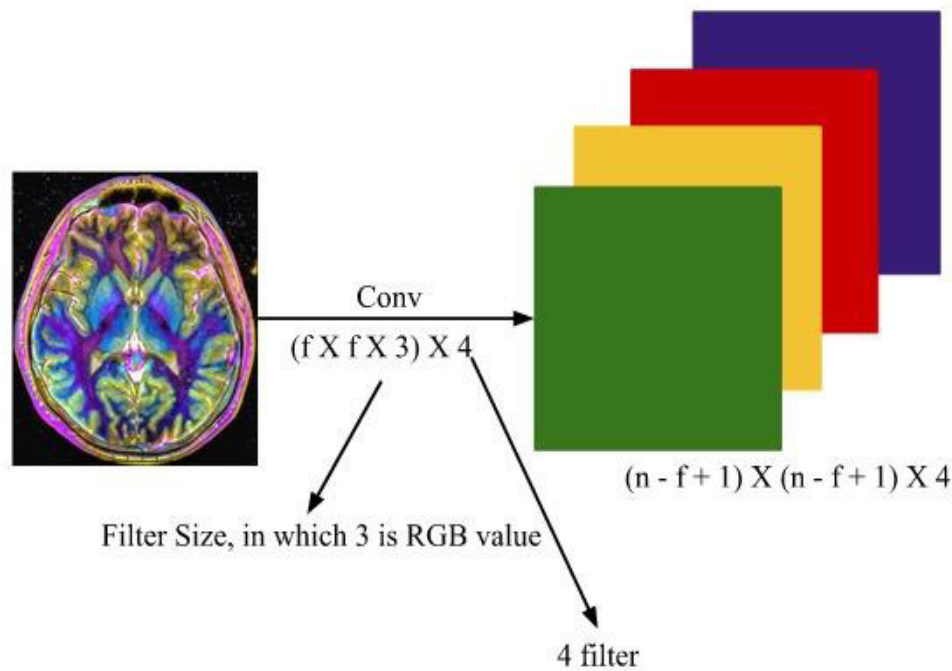


Fig: 23. Overall Convolution layer using non linear function and bias with multiple output in one frame

Let us see one image of MRI Brain as input which has the size 32×32 pixel with RGB value 3. So the image size is $32 \times 32 \times 3$. First we will convolve the input image with the particular filter size $5 \times 5 \times 3$ and use stride 1 with 4 filters so the output size will be $28 \times 28 \times 4$. Here the 28 is $n - f + 1 = 28$. As we use 4 numbers of filters we will get 4 dimensional output. After convolving this image we pass this to a max pooling layer. So here we are using max pooling 1 with the filter size 2×2 and stride 2. Thus the dimension image will become half, which means 28×28 will become $14 \times 14 \times 4$ and the number of channels will remain the same. In the traditional CNN both convolutional layer and max pooling layer will be considered in one layer. So, now this output $14 \times 14 \times 4$ will pass another convolutional layer which we name Conv2. Here we will use another filter size 3×3 but the 4 will remain the same. In the Conv 2 layer we used a filter so the output has 8 channels, again it passes through the max pooling layer and we will get the output half of the input image size. In this we consider layer 2. So, we can continue these layers as many times as we want. It depends on the type of application. If we are performing a completed application like healthcare then we might need our complex or large convolutional architecture. Also

the another thing is that we can completely skip this max pooling operation if we are building big CNN architecture because maxpooling will reduced the size of images that we are dealing with, and we might not want to reduce the size of images that we are dealing with, and we might not want to reduce size too much.

So once we are done with all convolutional layers and max pooling layers then it's time to add fully connected layers, before we apply these connected layers we need to flatten the final output image. So, the $6 \times 6 \times 8$ will be flatten only into the 1D factor which is 288 units in one flatten. Now once flattened it is now connected with a fully connected layer. This will represent FC3(Fully Connected layer), which indicates it as 3 layers in our CNN architecture. Let we keep 120 neurons in FC3 so it will connect with all nodes of flatten output. So the number of weight parameters is 288×120 . Now we add multiple fully connected layers as much as we want but, it is not fair adding fully connected layers highly increases the amount of parameters we need to deal with. Then in the final layer we will apply the sigmoid or softmax activation function depending on the types of applications we are making. For example if we are using binary classification then we will be using sigmoid or we will be using softmax activation. So, the final layer is called sigmoid or softmax or FC4 layer. In the softmax layer, there are huge numbers of neurons that must correspond to the number of output categories. This final layer is responsible for generating the predicted value. Prior to making predictions, it is imperative to train the system to ensure the factors are appropriately adjusted. Subsequently, the final predictor is utilized to compute the cost function, which quantifies the extent of inaccuracies in the model's predictions.

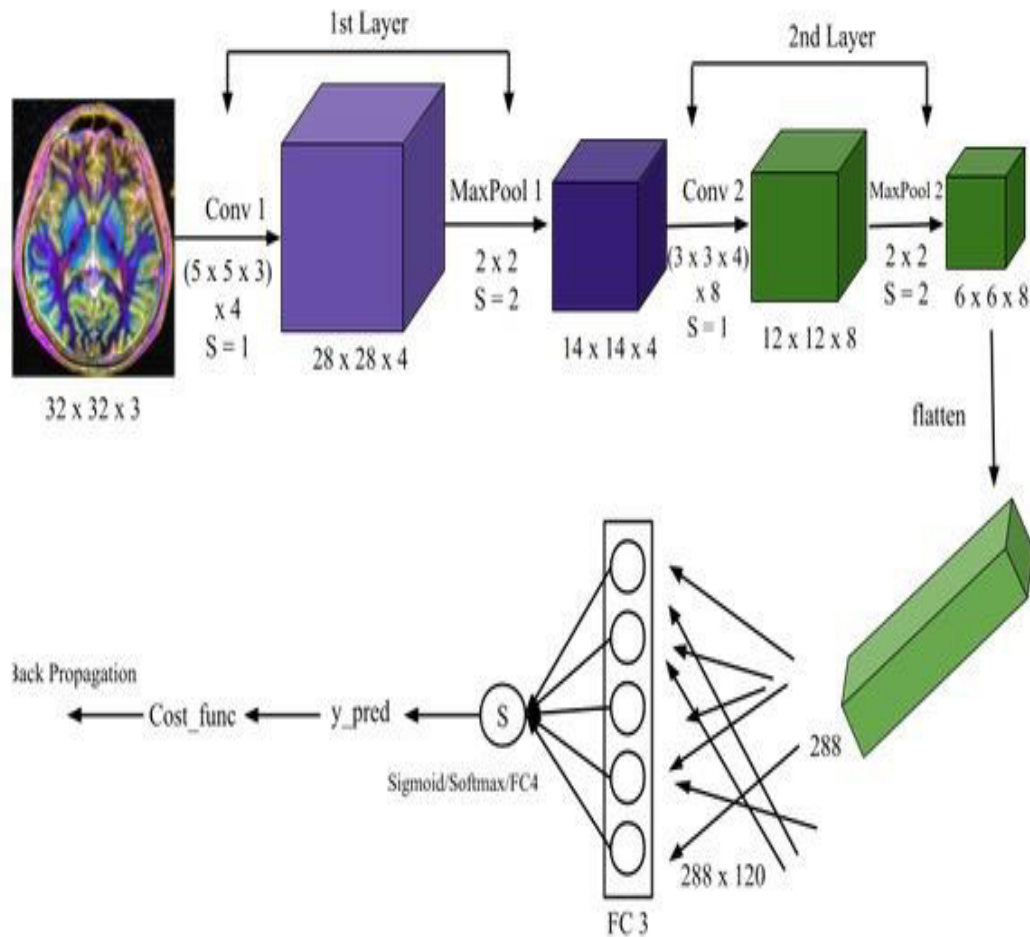


Fig. 24. The Architecture of Convolutional Neural Network

Now the question is which cost function is used? The answer is if we are using binary classification then it will be 'Bn' binary entropy cost function. If we are performing multiclass classification then we can use categorical class entropy classification. The goal is to lower the cost function so that we can train and improve our model's accuracy. So, how do we conduct the training? We employed the same old back propagation model as in our Artificial Neural Network or Neural Network.

In Artificial Neural Networks we use dense networks. There is lots of redundancy. The parameters are very high but we can reduce the amount of factors by using convolutional and max pooling. So we can reduce the size by convolutional layer and max pooling layer. The implementation of the back propagation algorithm in CNN is actually very complicated because here we also need to calculate the $\delta \text{ cost} / \delta w$. The

challenge is that the ‘w’ parameters are actually inside the filters used for the convolution operation.

Thus calculating this $\frac{\delta cost}{\delta w}$ and applying the gradient descent is very complicated in CNN. To make this job easy we use the Tensorflow framework in python. It makes our implementation very quick and easy. We do not need to deal with the details in the integrity of the complex technique but we can just mention the name. Example just mention what cost function we need or the name of back propagation we want to use Adam, or Momentum or gradient descent. So, tensorflow automatically implements for us. The advantage is to make CNN and find that the accuracy is not pretty good so you may want to add or remove some layer or modify the network. Whenever we modify the network layer we also need to do the same modification in back propagation as well. There are chances to lead to errors while doing back propagation. Thus all these will be very complicated and very time consuming and we won't be able to focus on the architecture properly but in tensorflow we just mention the type of layer like conv layer and parameters like stride, padding etc. If we want to use max pooling then mention it in there and so on.

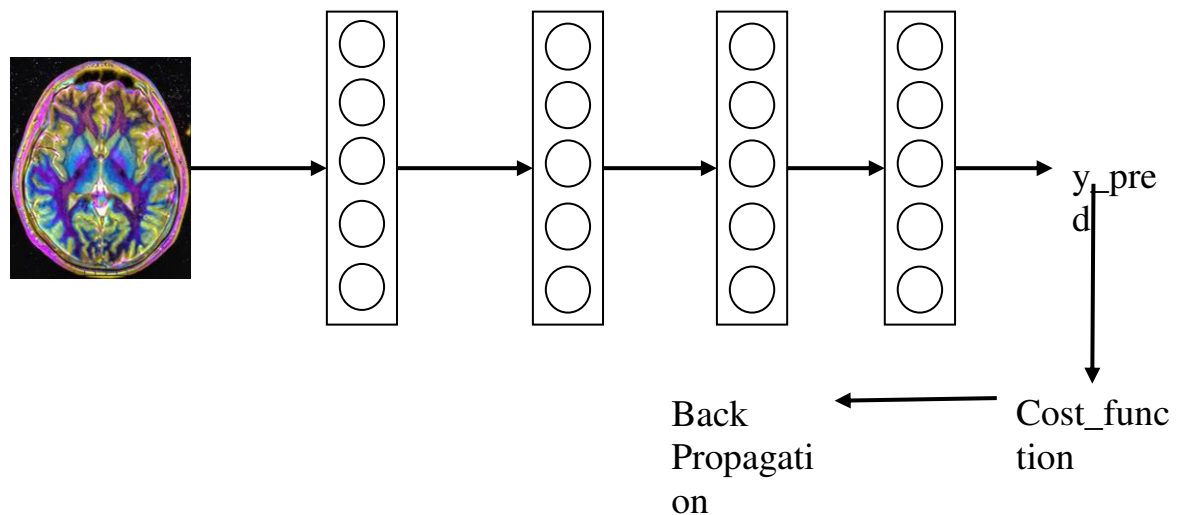


Fig: 25 : The Traditional Architecture of Artificial Neural Network

3.3.4 Deep Belief Network model, the feature extraction of skeletal image

A Deep Belief Network (DBN) is a form of artificial neural network made up of several layers of stochastic, latent variables known as hidden units. DBNs function as

generative models and are Well-suited for unsupervised learning tasks such as feature learning, dimensionality reduction, and data generation.

A DBN typically consists of several layers of neurons systematized in a hierarchical pattern. The network's initial layer is the visible layer, and it interacts directly with input data. The following layers are made up of hidden layers, with each layer having a unique number of neurons. DBNs frequently feature a symmetric architecture, with connections between each layer and its neighboring layers but not inside the same layer. Restricted Boltzmann Machines (RBMs) are probabilistic, generative neural networks that serve as the foundation of Deep Belief Networks. RBMs comprise two layers: visible and hidden, connected by symmetric connections. The architecture follows a bipartite graph structure, where neurons in the visible layer connect to neurons in the hidden layer, with no connections within each layer. RBMs adopt a probabilistic approach, employing binary units to represent the presence or absence of features. During training, RBMs learn to reconstruct input data by altering connection weights with approaches like Contrastive Divergence (CD) and Persistent Contrastive Divergence (PCD). DBNs are often trained using a greedy layer-wise strategy, which involves training each layer of the network independently before fine-tuning the entire network. The first layer (visible layer) is learned directly on the input data using unsupervised learning techniques like RBMs. After the first layer is trained, its output is utilized as input to train the next layer, and so on until all layers are trained. After training each layer independently, the complete network is fine-tuned using supervised learning techniques like backpropagation. Fine-tuning changes the weights of connections across the network to improve performance on a specific job, such as classification or regression. DBNs have been effectively used for a variety of applications, including image identification, audio recognition, natural language processing, and collaborative filtering. They are especially good at learning hierarchical representations of complex data and extracting significant features from high-dimensional input spaces. In general, a Deep Belief Network is a hierarchical, generative neural network design made up of numerous layers of RB Machine. Using a greedy layer-wise training technique, Deep BNs may learn hierarchical data representations and extract valuable features for a variety of machine learning tasks.

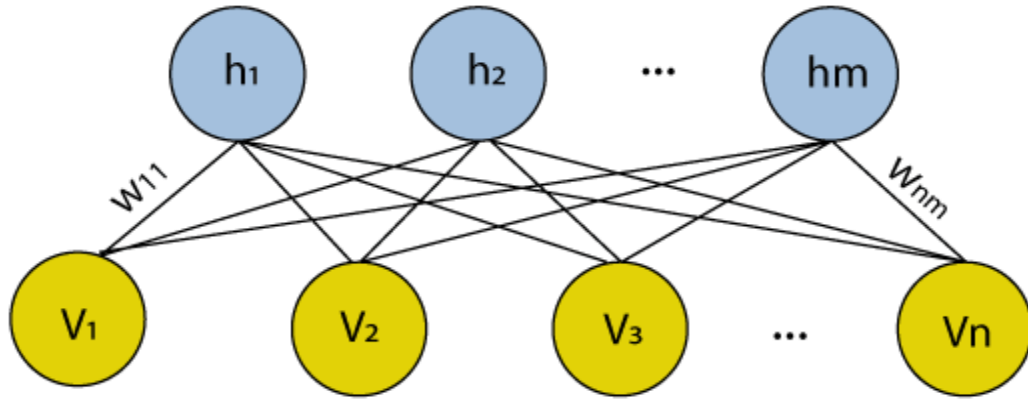


Fig 26 : Two-layer probabilistic neural network, RBM Architecture

Deep Belief Networks (DBNs) can be used in visual question answering (VQA) systems to extract meaningful features from both the visual and textual modalities of input data. In a VQA system, the input consists of both images and textual questions. The initial phase involves extracting features from both modalities. Regarding the visual modality, the DBN can be used to extract hierarchical representations of image features. Each DBN layer learns more abstract and complicated features from the image's raw pixel values. Similarly, for the textual modality, the DBN can process the text input (questions) to learn hierarchical representations of word embeddings or semantic features. Once features are extracted from both modalities, they must be joined or fused to form a single representation that captures the relationship between the image and the question. The extracted features from the visual and textual modalities can be concatenated, merged, or processed using additional layers of neural networks to create a joint representation. The composite representation should include relevant inputs derived from both the image and the question, allowing the VQA system to generate precise predictions. The DBN-based VQA system is trained using a massive dataset of matched images and questions, as well as their responses. During training, the DBN layer parameters, as well as any additional layers used for fusion and prediction, are optimized using backpropagation and stochastic gradient descent algorithms. The goal is to reduce the prediction error between the forecasted and ground-truth responses in the training data. Once trained, the DBN-based VQA system can be used to answer questions about unseen images. Given a new image and a question, the system first extracts features from both modalities using the trained

DBN. The retrieved characteristics are then integrated to form a joint representation, which is used by the prediction layer to generate the answer. The projected answer is often the result of a softmax layer, which represents the probability distribution of probable replies.

The performance of the DBN-based VQA system is examined using performance metrics such as accuracy, precision, recall, and F1 score. The system's ability to correctly answer questions about unseen images is assessed using a separate test dataset, and its performance is compared to other VQA systems using benchmark datasets. DBNs can be integrated into VQA systems to extract the features from both skeletal image and question-answer modalities, which are then fused to generate a joint representation for answering questions about images. By leveraging hierarchical feature learning, DBNs can capture complex relationships between visual and text, leading to improved performance in visual question-answering tasks.

A neural network of beliefs is a deep learning model made up of numerous layers of probabilistic latent variables that are often placed in a stack of restricted Boltzmann machines (RBMs). A DBM design consists of alternating levels of visible and hidden units, with each layer fully connected to the next. An RBM is a generative probabilistic neural network that represents the joint probability distribution of visible and hidden units. On an RBM, each visible unit is linked to every hidden unit, Nevertheless, there are no inter-unit connections within the same layer.. The connections between visible and hidden units are weighted, and each connection has its own weight parameter. RB Machines (RBMs) comprise a visible layer that represents input data, such as pixel values in an image, and one or more hidden layers that capture latent features or representations of the input. Deep Belief Networks (DBNs) are constructed by stacking multiple RBMs, creating a deep architecture. Each RBM's visible layer is connected to the hidden layer of the subsequent RBM, forming a feedforward chain. This enables the network to acquire hierarchical representations of the input data. Typically, DBNs are trained using layer-wise pretraining followed by fine-tuning via backpropagation. The pretraining step in a DBN initializes the weights of the RBMs layer by layer, allowing each layer to acquire informative data representations before fine-tuning the entire network. After layer-wise pretraining, the complete DBN undergoes fine-tuning using supervised

learning techniques such as backpropagation. This process involves training the DBN on a labeled dataset using gradient-based optimization algorithms to minimize a predetermined loss function, such as cross-entropy loss. Fine-tuning adjusts the weights of the entire network, enhancing its capability for classification or data generation. Each unit, whether visible or hidden, typically applies an activation function to its input to produce an output. The logistic sigmoid function, hyperbolic tangent function, and rectified linear unit (ReLU) function are all examples of common activation functions used in deep neural networks. The output layer of a DBN is determined by the individual task being handled. For example, in a classification assignment, the output layer could be made up of softmax units that represent different classes.

Overall, the architecture of a DB network consists of stacked RBMs with alternating layers of visible and hidden units. By learning hierarchical representations of leveraging the input data, DBNs have the capacity to apprehend complex patterns and dependencies, thereby proving effective in tasks such as feature acquisition, categorization, and generation.

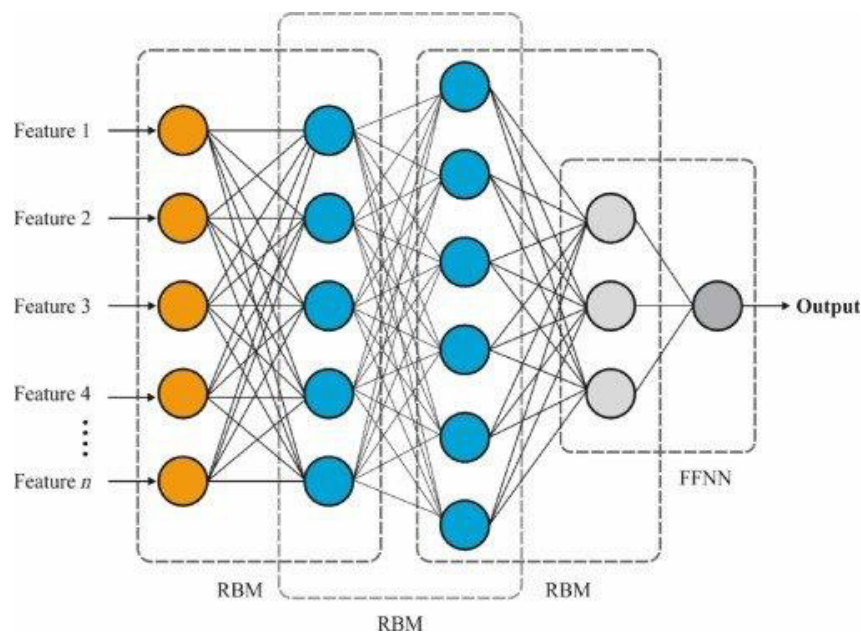


Fig: 27 : Overall architecture of a Deep Belief Network

A restricted Boltzmann machine is trained in a fundamentally different way from neural networks, which use stochastic gradient descent. There are two main steps to

train our dataset: the first one is Gibbs sampling, and the other one is the Contrastive Divergence Step.

Gibbs collection is the initial stage of training. When we are provided an input vector, we use the following $p(h|v)$ to forecast the hidden values h . However, if we know the hidden values h , we may utilize $p(v|h)$ to forecast the new input values.

$$p(v_i = 1 | h) = \frac{1}{1+e^{-(a_i+w_i h_j)}} = o'(a_i + \sum_j h_j w_{ij}) \quad (1)$$

This action is performed k times, generating a new input vector v_k that is derived from the initial input value v_0 .

$$p(v_i = 1 | v) = \frac{1}{1+e^{-(b_i+w_i v_j)}} = \sigma(b_j + \sum_i v_j w_{ij}) \quad (2)$$

The contrastive divergence stage modifies the weight matrix. The vectors v_0 and v_k are utilized to compute the activation probability of the hidden variables h_0 and h_k .

$$p(v_i = 1 | h) = \frac{1}{1+e^{-(a_i+w_i h_j)}} = o'(a_i + \sum_j h_j w_{ij}) \quad (3)$$

The update matrix is calculated as the difference between the outer products of the probabilities with input vectors v_0 and v_k , as shown in the matrix below.

$$\Delta W = v_o \otimes P(h_o|v_o) - v_k \otimes P(h_k|v_k) - v_k \quad (4)$$

We can now utilize this updated weight matrix to calculate new weight using gradient descent, as stated in the equation below.

$$W_{new} = W_{old} + \Delta W \quad (5)$$

To apply the architecture of a DB Network (DBN) to the task of visual question answering (VQA) using medical images, we need to adapt the network to handle both visual and textual input. The input layer of the DBN will consist of two components:

Visual input: This component represents the medical images. Each image is represented as a vector of pixel values or a higher-level feature representation extracted using techniques such as convolutional neural networks (CNNs).

Textual input: This component represents the textual questions associated with each image. Each question is represented as a vector using techniques such as word embeddings or one-hot encoding.

The DBN architecture will consist of alternating layers of visible and hidden units, similar to the traditional DBN architecture. However, each visible layer will now have two components: one for visual input and one for textual input. Each RBM in the stack will learn joint representations of the visual and textual inputs, capturing the relationships between the image content and the corresponding questions. During the layer-wise pretraining phase, each RBM in the stack will be trained independently using both visual and textual input. The pretraining phase initializes the weights of the RBMs, allowing them to learn hierarchical representations of both visual and textual features. After layer-wise pretraining, the entire DBN will be fine-tuned using supervised learning techniques. The network will undergo training using a dataset comprising labeled pairs of images and corresponding questions, aiming to predict accurate answers for each question based on the associated image. The output layer of the DBN will consist of units representing potential responses to inquiries, potentially employing a softmax activation function to produce probability distributions across potential answers. Integration of visual and linguistic components throughout the network will facilitate capturing the relationships between image content and textual queries. Through the hidden layers, the DBN will acquire composite representations of visual and linguistic attributes, enabling comprehension of the correlations between questions and images. By adapting the architecture of a DBN in this way, we can leverage its ability to learn hierarchical representations of complex data to effectively tackle the task of visual question answering in the medical domain. The network will be capable of understanding both the visual content of medical images and the textual queries associated with them, enabling it to provide accurate answers to a wide range of medical questions.

3.3.5 The deep insights of Region-based Convolutional Neural Network (R-CNN) for visual classification

The Region-based CNN (R-CNN) represents a seminal advancement in object recognition methodology, seamlessly integrating deep learning with conventional computer vision methodologies. Introduced by Ross Girshick et al. in 2014, it has

emerged as a foundational model in the realm of object detection. R-CNN, along with its subsequent iterations, has laid the groundwork for contemporary object detection architectures such as Faster R-CNN and Mask R-CNN. Unlike traditional object detection methods which relied on handcrafted features and disparate algorithms for localization and classification, R-CNN endeavors to unify these tasks within a unified deep learning framework, facilitating end-to-end training and enhanced performance. R-CNN begins by generating region recommendations via a selective search method. This algorithm looks for probable object regions in an image using color, texture, and other low-level information. Each region suggestion is then sent into a pre-trained CNN, such as AlexNet or VGG-16, which produces fixed-length feature vectors. The collected characteristics are then utilized to train a linear SVM classifier for object categorization. Each SVM is taught to distinguish between various object categories (for example, person, automobile, and dog). R-CNN also adds a bounding box regression step to refine the location of detected objects. This phase teaches how to change the bounding box coordinates predicted by the selective search method. R-CNN employs selective search to generate boundary recommendations from an user's input medical image. Each region suggestion is warped to a specific size and sent into a pre-trained CNN to extract features. The features are then sent via distinct SVM classifiers for object classification. Finally, the bounding box regression step adjusts the predicted bounding boxes to improve localization accuracy.

- R-CNN is trained in multiple stages:
 - Pre-training: The CNN is trained in advance on an extensive dataset. (e.g., ImageNet) for feature extraction.
 - Fine-tuning: The CNN is adapted on the target dataset for object detection.
 - SVM training: Separate SVM classifiers are trained for each object category using the extracted features.
 - Bounding box regression: Another model is trained to learn the adjustment needed for bounding box coordinates.

R-CNN has several limitations, including its slow inference speed due to the sequential processing of region proposals and the need to extract features for each proposal independently. This inefficiency makes it impractical for real-time

applications. Another limitation is the difficulty of end-to-end training, as the SVM classifiers and bounding box regressors are trained independently of the CNN. Several variants of R-CNN have been offered to address its limitations, including Fast R-CNN, Faster R-CNN, and Mask R-CNN. These variations strive to increase object detection speed and accuracy by using innovations like region proposal networks (RPNs) and shared feature extraction.

R-CNN takes a skeletal image as input and detects and localizes items in it. R-CNN starts by creating region proposals, which are candidate bounding boxes that may contain objects. These recommendations are often created with selective search, a standard computer vision technique that finds regions of interest based on color, texture, and other low-level properties. Selective search generates a huge number of area recommendations with different scales and aspect ratios. Each region proposal is cut from the input image and scaled to a predefined size, which typically corresponds to the input size predicted by a pre-trained convolutional neural network (CNN). The pre-trained CNN functions as a feature extractor and is often loaded with weights learned from a large-scale image classification challenge like ImageNet. The region proposals are fed through the CNN to produce fixed-length feature vectors. These feature vectors carry detailed semantic information about the content of each region proposal. The extracted feature vectors are subsequently fed into a sequence of fully connected layers, followed by a softmax layer for object categorization. Each completely linked layer serves as a classifier for a certain item category (such as a human, automobile, or dog). The softmax layer generates a probability distribution over the object categories, showing the likelihood that each region suggestion belongs to a specific category. In addition to object classification, R-CNN incorporates a bounding box regression phase to improve the localization of discovered objects. This phase teaches how to alter the coordinates of the bounding boxes predicted by the region proposal algorithm, improving their accuracy. Bounding box regression is often accomplished using a separate set of fully connected layers that forecast the offsets required to change the bounding box positions. The ultimate output of R-CNN comprises identified objects, their corresponding bounding boxes, and class labels. Each detected object is associated with a bounding box that precisely delineates its location within the input image, alongside a class label denoting its category.

- Training Stages of R-CNN in various steps:
 - Pre-training: The Convolutional NN is pre-trained on a large dataset (e.g., ImageNet) for feature extraction.
 - Fine-tuning: The Convolutional NN is optimized on the target dataset for object detection.
 - Object classification and bounding box regression: The fully connected layers in charge of object categorization and bounding box regression are trained by backpropagation and gradient descent.

R-CNN has inspired several variants, including Fast R-CNN, Faster R-CNN, and Mask R-CNN, which enhance upon its speed and accuracy by introducing innovations such as region proposal networks (RPNs) and shared feature extraction.

Region-based Convolutional Neural Network (R-CNN) can be used in a Visual Question Answering (VQA) system for image feature extraction by first selecting regions of interest within the image, and then extract features from these regions to answer image-related questions.

R-CNN starts by generating region proposals within the input image. These region proposals are possible bounding boxes that may contain objects of interest. These region suggestions can be generated using selective search or another region proposal algorithm that takes into account low-level image properties like color, texture, and shape. Once the region proposals have been developed, R-CNN crops and warps each one from the original image to a defined size. The cropped regions are then fed into a pre-trained CNN, such as VGG, ResNet, or Inception, to extract features. The CNN functions as a feature extractor and is often trained on a large-scale picture dataset such as ImageNet. The output of the CNN is a fixed-length feature vector that encodes rich semantic information about the content within each region proposal. In parallel to extracting image features, the input question is processed using natural language processing (NLP) techniques to standardize it into a fixed-length vector representation, known as a question embedding. Techniques like word embedding (e.g., Word2Vec, GloVe) and recurrent neural networks (RNNs) or transformer models (e.g., BERT) can be used to encode the question into a numerical vector. The feature vector obtained from the image regions and the question embedding are then fused or concatenated to create a joint representation that combines both visual and

textual information. Various fusion methods can be employed, such as element-wise addition, concatenation, or attention mechanisms, to successfully integrate image characteristics and question embeddings effectively. The joint representation is then passed through a classifier, such as a multi-layer perceptron (MLP) or a softmax layer, to predict the answer to the input question. The classifier learns to map the joint representation to a probability distribution over a predefined set of answer options. The entire VQA system, including the R-CNN for image feature extraction, is trained end-to-end using a large dataset of image-question-answer triplets. During training, the parameters of the CNN, question embedding model, fusion method, and classifier are optimized using backpropagation and gradient descent to minimize the prediction error. During inference, given a new image and question, the trained VQA system predicts the most likely answer by passing the image through the R-CNN for feature extraction, processing the question to obtain its embedding, fusing the image features and question embedding, and finally predicting the answer using the trained classifier.

By leveraging R-CNN for image feature extraction in a VQA system, the model can effectively extract visual features from specific regions of interest within the radiology image, enabling it to accurately answer questions about the content of the image.

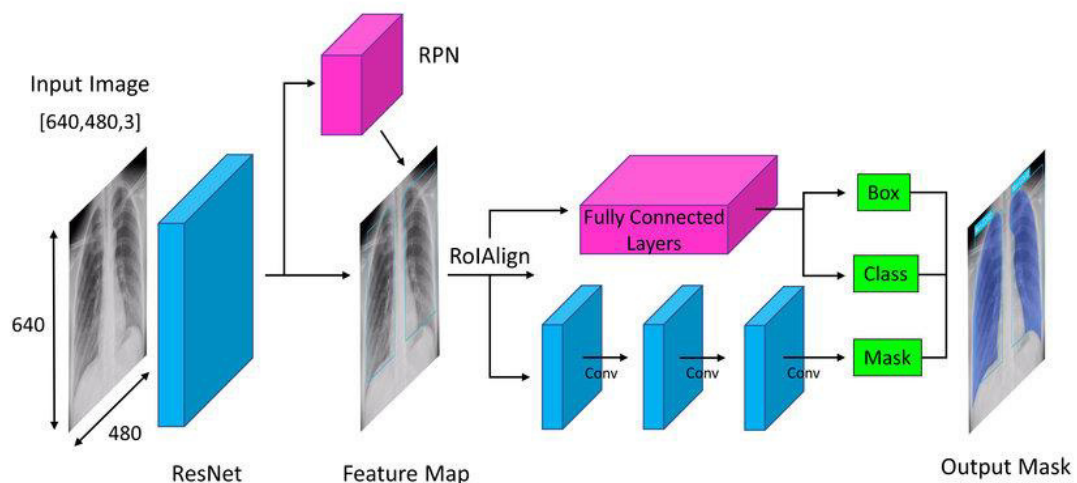


Fig: 28: Faster Region-based Convolutional Neural Network (R-CNN)

Faster R-CNN enhances the original R-CNN by integrating a Region Proposal Network (RPN) directly into the architecture. This integration facilitates end-to-end training, leading to improved efficiency and accuracy in object detection tasks. The

RPN operates as a fully convolutional network, leveraging feature maps extracted from input images to generate region proposals, which represent potential bounding boxes containing objects. By sliding a tiny network across the feature map, the RPN anticipates the presence of objects ("foreground") within each anchor box and refines their positions. The Region Proposal Network (RPN) in Faster R-CNN generates region proposals by integrating expected bounding box coordinates with anchor boxes of different scales and aspect ratios. This process aims to efficiently cover potential object locations across the image. Additionally, Faster R-CNN utilizes a deep convolutional neural network (CNN) as a feature extraction backbone to extract high-level features from the input image. Commonly used backbone networks include VGG, ResNet, and MobileNet, which are pretrained on extensive image classification datasets like ImageNet. These backbone networks process the input image to produce a feature map containing detailed spatial information. After generating region proposals, Faster R-CNN uses RoI pooling to extract fixed-size feature vectors from each proposal. RoI pooling partitions each proposed region into a predetermined number of spatial bins, subsequently performing max pooling within each bin to yield a feature vector of fixed dimensions.

This ensures that the features extracted from different-sized region proposals have the same spatial dimensions, which is essential for feeding them into subsequent layers of the network. RoI pooling generates fixed-size feature vectors, which are used in fully connected layers for object classification and bounding box regression. The classification branch assigns a class label to each proposal based on its likelihood to contain an item. The regression branch refines the coordinates of the bounding box to better localize the object within the proposal. Both the classification and regression branches are typically implemented as fully connected layers on top of the shared feature representation. After generating region proposals, Faster R-CNN employs Region of Interest (RoI) pooling to derive fixed-size feature vectors from individual proposals. This process entails segmenting each area proposal into a predefined number of spatial bins and subsequently applying max pooling within each bin to produce a fixed-size feature vector. This process ensures that features extracted from region proposals of different sizes have consistent spatial dimensions, which is necessary for subsequent layers in the network. The resulting fixed-size feature vectors

are then input into fully connected layers for object classification and bounding box regression. The classification branch determines the likelihood that each proposal contains an object and assigns a class label.

Faster R-CNN can help with feature extraction for visual or skeletal image question answering (VQA) in radiology images by efficiently detecting and localizing relevant objects or regions within the images. Faster R-CNN can accurately localize and identify regions of interest (ROIs) within radiology images that contain anatomical structures, abnormalities, or other medically significant features. These ROIs can serve as the basis for answering questions related to specific areas or abnormalities within the image. Once the relevant regions are identified, Faster R-CNN can extract high-level features from these regions using RoI pooling. This process involves pooling characteristics from the convolutional layers of the network within the identified regions, allowing for the extraction of rich and discriminative feature representations. By detecting and localizing objects within radiology images, Faster R-CNN helps in understanding the semantic content of the images. This semantic understanding can be crucial for answering questions that require interpretation of the visual content, such as identifying anatomical structures, pathologies, or medical devices.

Faster R-CNN offers cutting-edge performance in object detection tasks, providing accurate localization of objects with high precision and recall. This improved accuracy ensures that the extracted features are relevant to the task at hand, thereby enhancing the performance of the VQA system. Faster R-CNN can be seamlessly integrated into the VQA pipeline, allowing for end-to-end training of the system. This integration enables joint optimization of the object detection and question answering components, leading to better alignment between the visual and textual modalities and improved overall performance.

Overall, Faster R-CNN serves as a powerful tool for feature extraction in VQA systems for radiology images, facilitating the identification of relevant visual content and enhancing the system's ability to provide accurate and informative answers to medical-related questions.

Multiple CNN-based object detection algorithms have been developed to address various challenges in computer vision tasks. These algorithms include R-CNN [87], SPPnet[88], Fast R-CNN [89], Faster R-CNN [90], You Only Look Once (YOLO) [91], and Single Shot Multibox Detector (SSD) [92]. Among these, Faster R-CNN stands out for its cutting-edge performance in object detection tasks.

Faster R-CNN is renowned for its accuracy and ability to precisely localize objects within images. While it may have a slightly slower processing speed compared to newer approaches like YOLO and SSD, its performance and versatility make it an ideal choice for many applications, including visual or skeletal image question answering in the healthcare domain.

One of the principal advantages of Faster R-CNN lies in its capacity to efficiently execute object identification tasks. It achieves this by employing a region proposal network (RPN) to create region recommendations directly from convolutional feature maps. These region proposals are then used to localize objects within the image, allowing for precise object detection.

Another important characteristic of Faster R-CNN is its flexibility in handling images of varying sizes. Unlike some other object detection algorithms that require fixed input image sizes, Faster R-CNN can efficiently process images of different sizes using a sliding window approach. This makes it particularly useful for dealing with large medical images commonly encountered in radiology.

Furthermore, Faster R-CNN does not compromise on accuracy despite its efficiency. It leverages deep convolutional neural networks (CNNs) to extract rich feature representations from images, enabling robust object detection and classification. Additionally, it incorporates advanced techniques such as region of interest (RoI) pooling to further enhance its performance.

Faster R-CNN consists of two main modules: the region proposal network (RPN) and the object detection network. The RPN is responsible for generating region proposals, while the object detection network classifies the objects within these proposed regions.

The workflow of Faster R-CNN begins with the radiology input image being processed by a feature extractor, typically a convolutional neural network (CNN) with

convolutional and pooling layers. This feature extractor generates feature maps, which contain high-level representations of the image that capture important visual information.

The feature maps extracted from the input image are subsequently fed into the region proposal network (RPN) which examines them to determine whether regions are likely to contain objects. The RPN achieves this by traversing a small window (typically 3x3) across the feature maps, predicting the presence of an object within each window, and generating bounding box proposals for these regions. This process allows the network to propose candidate regions of interest for further analysis and classification.

Once the region proposals are generated by the RPN, they are passed to the object detection network, which further processes each region to classify the objects within them and refine their bounding box coordinates if necessary. This network typically consists of additional convolutional and fully connected layers, followed by classification and regression heads.

In the original Faster R-CNN architecture, the feature extractor often relies on established CNN architectures such as VGG-16. However, alternative models may offer better performance. For instance, Inception-ResNet, which merges the Inception and ResNet architectures, surpasses VGG-16 in certain applications.

The region proposal network (RPN) of Faster R-CNN examines feature maps using a sliding window approach. However, utilizing a fixed-size window could be limiting, especially when dealing with objects of diverse shapes and sizes. Faster R-CNN mitigates this concern by introducing the notion of anchor boxes.

Anchor boxes are predetermined bounding boxes with diverse scales and aspect ratios that are distributed across the image. These anchor boxes provide reference points for spotting items of various shapes and sizes. By using multiple anchor boxes, the RPN can better capture the variability in object shapes within the image.

During the sliding window operation, each anchor box generates two scores: one indicating whether the window contains an object (foreground) or not (background), and the other indicating the class of the object if it is present. Specifically, for each anchor box, the RPN produces $4 \times k$ pieces of information corresponding to the

bounding box coordinates (i.e., the coordinates of the box's top-left corner and bottom-right corner) and $2 \times k$ pieces of information corresponding to the objectness score and the class score.

Using anchor boxes with multiple scales and aspect ratios, the RPN can successfully handle objects of various forms and sizes, improving object detection accuracy in Faster R-CNN. This approach allows Faster R-CNN to perform consistently over a wide range of object types and combinations in input images. R-CNN's region proposal network (RPN) produces $4k \times W \times H$ pieces of area information and $2k \times W \times H$ pieces of class information, where k is the number of anchor boxes used and $W \times H$ is the size of the feature map.

Initially, Faster R-CNN typically utilizes 9 anchor boxes ($k = 9$), each with a combination of scales and aspect ratios. However, to improve the detection of specific types of objects, such as glomeruli in medical images, the number of anchor boxes can be increased. In this case, 12 anchor boxes ($k = 12$) were employed, including variations in scale and aspect ratio to better capture the diversity of object shapes and sizes.

Since the RPN may propose multiple candidate regions for the same object, it introduces redundancy in the detection process. To address this issue, non-maximum suppression (NMS) is applied based on the class scores associated with the proposed regions. NMS is a technique used to remove redundant bounding boxes by selecting the most confident detection and discarding overlapping detections with lower confidence scores. By applying NMS, the number of proposed regions of interest (ROIs) is typically reduced to a more manageable number, typically around 300, ensuring that only the most relevant and confident detections are retained for further processing.

In Faster R-CNN, the region of interest (ROI) pooling technique is employed to convert ROIs of varying sizes into fixed-sized feature vectors. ROI pooling ensures that the input to subsequent layers remains consistent in size, regardless of the size or shape of the original ROIs. This process is crucial for maintaining spatial information while enabling the network to handle inputs of different dimensions.

Following ROI pooling [87], the fixed-sized feature vectors are sent through a fully linked layer. This layer is responsible for two major tasks: bounding box regression and object classification.

Bounding box regression predicts the coordinates of the bounding boxes that surround the identified items. These coordinates usually contain the upper-left and lower-right corners of the bounding box.

Object categorization seeks to assign a class label to every identified object. In the case of Faster R-CNN, the classification problem is distinguishing between different object classes, with an additional class for the background. The network is trained to classify objects into $n + 1$ classes, where n represents the number of distinct object categories, and the additional class represents the background.

Both bounding box refinement and object classification are supervised tasks, meaning that the network's predictions are compared to ground truth annotations during training to optimize the network parameters and improve its performance.

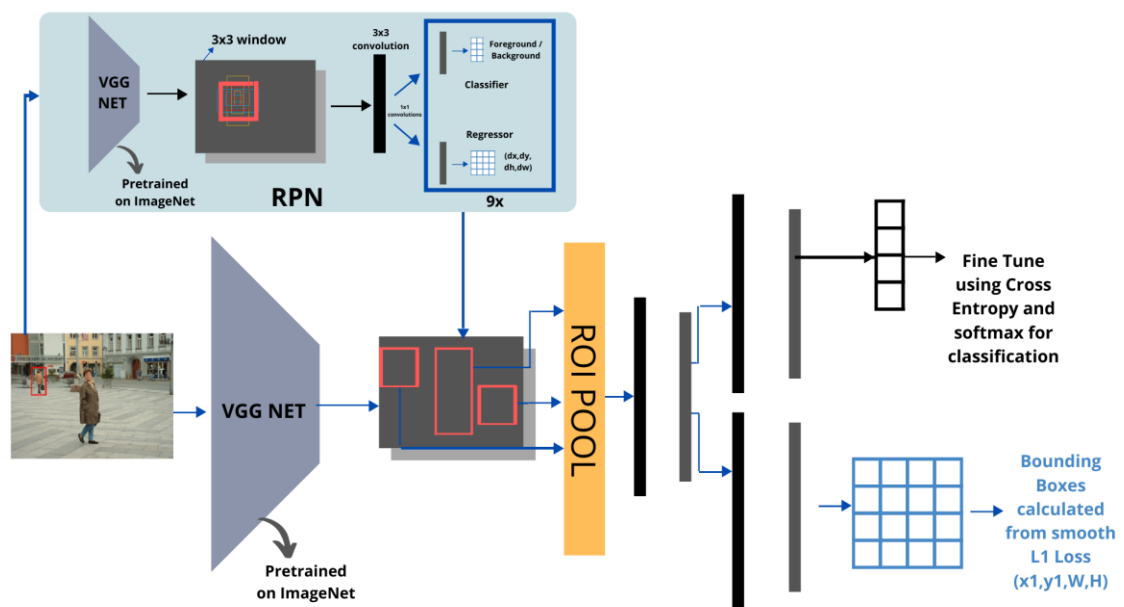


Fig 29 : Understanding the Fast RCNN and Faster RCNN Architecture

3.3.7 Long Short-Term Memory networks for question answering system

Long Short-Term Memory Networks (LSTMs) are a sort of a recurrent neural network (RNN) architecture developed to solve the vanishing gradient problem., which impedes the training of standard RNNs on long sequences. LSTMs have

emerged as powerful tools for modeling sequences, with applications in natural language processing, speech recognition, and time series analysis. LSTMs are designed to capture long-term dependencies in sequential data while limiting the influence of disappearing gradients. They use a memory cell and a set of gates to regulate information flow, allowing them to selectively keep or discard information over time. The memory cell is the central component of an LSTM, temporally persistent storage. The memory cell has a self-connected recurrent connection, allowing it to maintain its state over multiple time steps. The cell state, denoted as C_t is updated at each time step based on the input, previous cell state, and gate activations.

- LSTMs use three types of gates to control the flow of information: input gate (i_t), forget gate (f_t), and output gate (o_t).
- Each gate is made up of a logistic function or the normal logistic function, which produces values ranging from 0 to 1 that indicate how much information should be allowed through.
- The input gate regulates the flow of new information into the cell state.
- The forget gate controls the extent to which previous information should be forgotten from the cell state.
- The output gate controls how much of the cell state is revealed to the network output.

The gates are computed using following equations

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (6)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (7)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (8)$$

σ represents the sigmoid activation function, W denotes weight matrices, h_{t-1} is the previous hidden state, and x_t is the current input.

The cell state is updated using following equations

$$\underline{C_t} = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (9)$$

$$C_t = f_i \cdot C_{t-1} + i_t \cdot \underline{C}_t \quad (10)$$

\underline{C}_t represents the new candidate values to be added to the cell state.

Finally, the output of the LSTM at time step t , denoted as h_t , is computed as:

$$h_t = o_t \cdot \tanh(C_t)$$

The output gate regulates the exposure of the cell state C_t to produce the final hidden state h_t for the current time step.

LSTMs are commonly trained using optimization techniques such as backpropagation through time (BPTT) or gradient descent, with gradients computed using backpropagation through time (BPTT) or more sophisticated methods like Long Short-Term Memory networks in deep learning. Throughout the training process, the LSTM's parameters, including gate weights and biases, are iteratively adjusted to minimize a specified loss function, such as cross-entropy loss or mean squared error.

Several strategies and versions of Long Short-Term Memory (LSTM) networks were created to address certain difficulties or suit particular use cases. The conventional LSTM design is made up of memory cells, input gates, forget gates, and output gates, as indicated in the previous detailed perspective. BiLSTMs analyze input sequences in both forward and backward orientations, allowing the network to collect data from both past and future contexts concurrently. This leads to a better understanding of context and enhanced performance in tasks such as sequence labeling and sentiment analysis.. The placed LSTMs are made up of many LSTM layers placed on top of one another. Each layer gets input from the preceding layer and produces output for the subsequent layer. Stacking LSTMs enables the model to acquire hierarchical representations of sequential data, facilitating the capture of intricate patterns and dependencies. Alternatively, GRUs offer a simpler architecture with fewer parameters compared to LSTMs. GRUs consolidate the functionalities of input and forget gates into a single "update gate," enhancing computational efficiency while preserving long-term dependencies. Additionally, attention mechanisms enhance LSTM networks by enabling them to direct attention towards specific segments of the input sequence during output generation. Attention mechanisms enable the model to adaptively allocate its focus to salient information, dynamically assigning varying degrees of importance to different segments of the input sequence. This dynamic

weighting facilitates improved performance across various tasks such as machine translation, image captioning, and summarization. Several LSTM variations have been created to solve distinct issues in different domains. For example, medical LSTM variants may incorporate domain-specific features or pre-training on medical data to improve performance in healthcare-related tasks. Researchers and practitioners often develop customized LSTM architectures tailored to specific tasks or datasets. These customized architectures may include additional layers, connections, or modifications to the standard LSTM architecture to optimize performance for the given task.

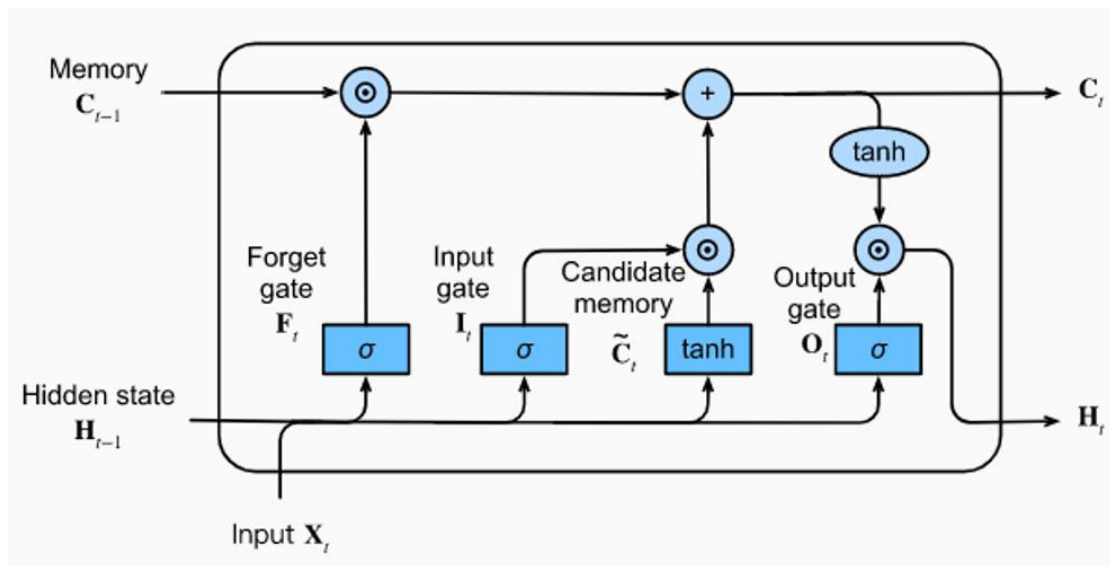


Fig 30 : Architecture of Long Short-Term Memory (LSTM) networks

3.4 The proposed Approach for Visual Question Answering System (VQAS) for skeletal Images

The Visual Textual System or VQA model represents a significant advancement in artificial intelligence research, aiming to equip machines with the ability to grasp visual content and effectively respond to questions posed about it. In the domain of medical visual question answering (Med-VQA), this technology takes on added importance as it endeavors to provide accurate responses to clinical queries based on radiological images.

In Med-VQA, the task involves presenting the computer with a radiological image alongside a relevant clinical question. The objective is to develop a sophisticated system capable of analyzing the visual information within the image and generating

appropriate answers to the posed questions. This demands a comprehensive understanding of both medical imaging and natural language processing, making Med-VQA a challenging yet promising field of study.

The ultimate goal of Med-VQA is to enhance medical diagnostics, decision-making, and patient care by leveraging the capabilities of artificial intelligence to assist healthcare professionals in interpreting radiological images and extracting valuable insights from them. By seamlessly integrating image analysis with question answering capabilities, Med-VQA has the capability to innovate medical imaging interpretation and make a substantial contribution to the advancement of healthcare practices.

Through rigorous research and development efforts in Med-VQA, we aim to achieve breakthroughs that not only augment the efficiency and accuracy of clinical decision-making but also enhance patient outcomes and overall healthcare delivery. With its formalized methodologies and interdisciplinary approach, Med-VQA represents a pivotal area of exploration at the intersection of artificial intelligence and healthcare, with far-reaching implications for the future of medical diagnostics and treatment.

3.4.1 Skeletal image Feature Extraction using Block_12_add Faster R-CNN (B12- FRCNN) algorithm

Block_12_add Faster R-CNN (B12-FRCNN) extends the Faster R-CNN (Region-based Convolutional Neural Network) approach, which is widely used for object detection. B12-FRCNN improves the performance of the Faster R-CNN architecture by including an additional block known as Block 12.

B12-FRCNN has two basic components: The region proposal network (RPN), as well as the object detection network. The RPN creates region proposals, which are candidate bounding boxes that may include objects of interest. The object detection network refines and classifies these proposals before producing the final detection findings.

B12-FRCNN inserts Block 12 into the Faster R-CNN network's feature extraction backbone. This block is often made up of numerous convolutional layers, followed by activation functions and pooling layers. Block 12's goal is to extract additional discriminative features from the input skeletal images, increasing the accuracy of object detection.

Block 12's design may change depending on the application and task requirements. It can be tailored to the complexity of the dataset, the diversity of the items to be detected, and the availability of computer resources.

Overall, B12-FRCNN extends the capabilities of the original Faster R-CNN algorithm by incorporating additional layers for feature extraction, thereby enhancing its ability to detect objects accurately and efficiently. This makes it a valuable tool for a wide range of computer vision applications, including object detection in images and videos, surveillance, autonomous driving, and medical imaging.

Block_12_add Faster R-CNN (B12-FRCNN) represents an enhanced iteration of the Faster R-CNN approach, renowned for its state-of-the-art capabilities in object detection within images. B12-FRCNN elevates the efficacy of Faster R-CNN by integrating a novel component known as Block 12 into its architecture. This pivotal block is seamlessly integrated into the feature extraction backbone of the network and is specifically designed to extract more discriminative features from input images. Faster R-CNN, a two-stage object detection methodology, comprises the region proposal network (RPN) and the object detection network. The RPN is responsible for generating region proposals, which serve as candidate bounding boxes potentially containing objects of interest. Subsequently, the object detection network refines and categorizes these proposals, ultimately yielding the definitive detection outcomes. B12-FRCNN adds an additional block, known as Block 12, to the Faster R-CNN network's feature extraction backbone. Block 12 is made up of many convolutional layers, subsequent to activation functions and pooling layers. The goal of Block 12 is to extract additional discriminative features from the input images, allowing the network to catch finer details and subtleties in the visual information. In the feature extraction process, the input image is transmitted via Block 12's convolutional layers. Each convolutional layer uses a sequence of learnable filters to extract features from the input image, including edges, textures, and forms. Activation functions, such as ReLU (Rectified Linear Unit), are then used to add nonlinearity to the network and allow it to learn complex patterns. Pooling layers play a crucial role in feature map downsampling, effectively reducing spatial dimensions while retaining essential information. In the case of Block 12, retrieved feature maps undergo integration into the Faster R-CNN framework. These enhanced feature maps serve as the foundation

for both the region proposal network (RPN) and the object detection network. The RPN utilizes these feature maps to generate region ideas, subsequently refined and categorized by the object detection network to discern objects within the image. Through the incorporation of Block 12 into the Faster R-CNN architecture, B12-FRCNN achieves notable enhancements in object detection accuracy and efficiency. The additional layers within Block 12 enable the network to extract more intricate and nuanced characteristics from input images, thereby augmenting detection performance. B12-FRCNN is particularly effective in scenarios where detecting small or intricate objects is challenging, as it can extract finer-grained features from the images.

In the proposed approach system, the feature extraction process plays an indispensable role in extracting discriminative features from radiology images to enable accurate question answering. Choosing the best feature extraction layer is critical for obtaining good performance in the visual question answering (VQA) assignment. Input data for the training phase includes radiological images and question-and-answer pairs. The goal is to train the VQA system to correctly answer questions concerning radiological images. The first stage of the training process involves extracting features from radiological images using the suggested Block_12_add Faster R-CNN (B12-FRCNN) method. B12-FRCNN is used to extract image characteristics because it can capture finer-grained visual information than the current Faster R-CNN method. In B12-FRCNN, the "block_12_add" layer is used as the visual feature extraction layer instead of the conventional "activation_40_relu" layer used in Faster R-CNN. This modification addresses the gradient vanishing problem associated with the "activation_40_relu" layer and improves the quality of the extracted features from medical images. The selection of the optimal feature extraction layer, such as "block_12_add," is a critical aspect of the training phase. Empirical analysis is conducted to evaluate the performance of different feature extraction layers and determine which layer yields the best results for the VQA task. This analysis involves training the VQA system using different feature extraction layers, including "block_12_add" and other candidate layers. Each configuration's performance is appraised using measures such as accuracy, precision, recall, and F1-score on a validation dataset. The layer that achieves the highest performance metrics

is selected as the optimal feature extraction layer for the VQA system. After selecting the optimal feature extraction layer, the VQA system is validated on a separate validation dataset to ensure its generalization performance. Fine-tuning may be performed to further optimize the VQA system's performance in accordance with the validation results. In the testing phase, the trained VQA system is evaluated on unseen radiology images and question-answer pairs to assess its performance in real-world scenarios. By empirically analyzing and selecting the optimal feature extraction layer, the proposed system ensures that the VQA system can effectively leverage visual information from radiology images to generate accurate answers to clinical questions.

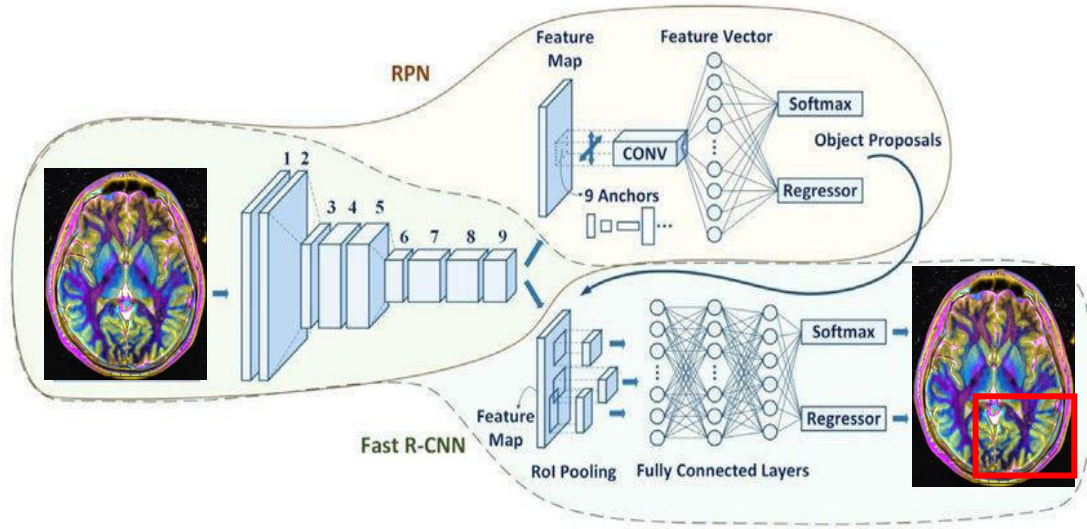


Fig 31 : B12-FRCNN, the "block_12_add" layer

B12-FRCNN begins by extracting features from medical images using a CNN backbone. Let's denote the input image as I , and the feature extraction function as $CNN(I)$. The output of this process is a set of feature maps denoted as $Feat = \{feat_1, feat_2, \dots, feat_N\}$, where each $feat_i$ represents a feature map produced by the CNN backbone.

The RPN generates candidate regions of interest (RoIs) by scanning the feature maps $Feat$. Let p_i denote the position and scale information of the i^{th} anchor box, and f_i denote the feature vector extracted from the corresponding region of the feature map. The RPN predicts the probability $p(object)$ and adjusts the coordinates p_i to refine the proposed bounding boxes. RoI pooling extracts fixed-size feature vectors from each RoI to facilitate subsequent processing. Let RoI_i denote the i^{th} RoI, and

$feat_{RoI_i}$, denote the feature vector extracted from RoI_i . The RoI pooling operation transforms variable-sized RoIs into fixed-size feature vectors by dividing the RoI into a grid and applying max pooling within each grid cell. The extracted RoI features are passed through classification and regression heads. Let f_{RoI_i} denote the feature vector of RoI_i after RoI pooling. The classification head predicts the probability distribution over object classes $p(class|RoI_i, Feat)$, while the regression head refines the bounding box coordinates p_i based on the extracted features $Feat$. Finally, B12-FRCNN integrates with a question answering (QA) module to answer clinical queries based on the detected objects and their contextual information. The QA module processes textual questions Q related to the medical images and utilizes the detected objects' features $Feat$ to generate appropriate answers. The integration of object detection and question answering can be represented as follows: $Answer = QA(Feat)$. This equation indicates that the answer generated by the QA module is a function of the textual question Q and the extracted features $Feat$ from the medical images.

Overall, B12-FRCNN combines advanced object detection techniques with question answering capabilities, leveraging both skeletal image and question answer information to provide accurate and meaningful responses to clinical queries related to medical images.

3.4.2 The proposed approach Kai-Bi-LSTM for Question Answering feature extraction

The Question-Answer (QA) pairs undergo preprocessing, starting with the splitting of the question text into individual words. Subsequently, stop word removal eliminates frequently occurring words like "a", "an", and "the", which do not contribute to the text's meaning. The root forms of words are then extracted via stemming, which condenses words with different suffixes under the same root word. This process minimizes the index size and improves computational efficiency.

After preprocessing, keywords are extracted from the question, representing highly relevant information. Keyword extraction entails obtaining relevant information from a text. For the response, text-only splitting is used. This step's output is in string

format, which is then translated to numerical format for effective categorization. This conversion is based on the word representation hypothesis.

LogishBERT is used for this purpose, a variation of Bidirectional Encoder Representations from Transformers (BERT) that employs a Logarithmic Swish activation function to handle complex data more effectively than the traditional GELU activation function. It is an activation function commonly used in deep learning models, significantly in neural networks. GELU is designed to approximate the Gaussian cumulative distribution function and has been found to perform well in various tasks, including natural language processing and computer vision. It is defined mathematically as:

$$GELU(x) = x \cdot P(X \leq x) = x \cdot \Phi(x) \quad (10)$$

where $\Phi(x)$ represents the cumulative distribution function of the standard normal distribution.

LogishBERT converts the output from its fully connected layer into numerical data, referred to as a score value.

The radiological image feature and LogishBERT score value are merged and used to train a Kai-Bi-LSTM classifier to predict answers. This classifier employs Bidirectional Long Short-Term Memory (BiLSTM) networks to address dependencies and different timelines. The Kaiming Initialization function is used to set the activation value for the forget gate.

During the testing phase, the system receives a sample image and question, initiating feature extraction and preprocessing procedures akin to those employed during training. The extracted features and corresponding scores are then fed into the classifier, which predicts a score value corresponding to the answer. Subsequently, this score value and the question are cross-referenced with the LogishBERT lexicon to determine the appropriate response.

In the results analysis section, the performance of the proposed system is assessed by juxtaposing it against existing methodologies, scrutinizing its innovative aspects, and evaluating various performance metrics including accuracy, precision, recall, and F-measure. Figure 31 illustrates a block diagram delineating the proposed technique.

LogishBERT is an adaptation of Bidirectional Encoder Representations from Transformers (BERT), a pre-trained Linguistic-focused model understanding tasks. BERT has gained significant attention in the field of textual feature extraction processing due to its effectiveness in capturing contextual information and semantic relationships in text data. LogishBERT enhances the traditional BERT model by introducing a logarithmic swish activation function, which aims to handle complex data more effectively. BERT, an acronym for the same, adopts a transformer-based methodology to glean contextual insights from input text. Its architecture comprises multiple layers of bidirectional encoders, facilitating the extraction of nuanced contextual information. BERT undergoes pre-training on extensive text corpora, engaging in two unsupervised learning tasks: masked language modeling (MLM) and next sentence prediction. Renowned for its exceptional performance, BERT has showcased state-of-the-art results across a spectrum of natural language processing tasks, encompassing text classification, named entity recognition, and question answering. The traditional BERT model uses the GELU (Gaussian Error Linear Unit) activation function in its fully connected layers. LogishBERT replaces the GELU activation function with a logarithmic swish activation function. The logarithmic swish function is defined as.

$$\text{logish}(h) = \ln(1 + e^x) \quad (11)$$

Compared to the GELU function, the logarithmic swish function introduces a logarithmic term in the activation function, which helps in handling complex data distributions more effectively. The logarithmic swish function is believed to provide smoother gradients and better performance on tasks with complex data distributions. After processing the input text data through the LogishBERT model, the output from the fully connected layers is converted into numerical scores. These numerical scores represent the likelihood or confidence of different answers or predictions. The scores are typically obtained by applying a softmax function to the output vector, which normalizes the values to probabilities, ensuring that they sum up to one. LogishBERT is particularly useful in question answering systems, where it can effectively process and understand textual data to generate accurate answers to user queries. It is integrated into the proposed system architecture described earlier, where it plays a

crucial role in converting textual input (questions and answers) into numerical representations for further processing and classification.

The Kaiming Initialization, also referred to as He Initialization, is a technique utilized to instigate the parameters of deep neural networks, including RN networks (RNNs) such as the Bidirectional LSTM (BiLSTM) model. This initialization method determines the initial weights of the neural network layers, encompassing the recurrent connections within the BiLSTM model. Its primary objective is to set the weights' initial values in a manner that prevents them from being excessively small or large, thereby mitigating issues related to vanishing or exploding gradients during the training process. Kaiming Initialization takes into consideration the activation function employed by the network, such as hyperbolic tangent (tanh) or rectified linear unit (ReLU), to ensure optimal performance. It adjusts the scale of initialization based on the characteristics of the activation function to ensure that the activations do not saturate too quickly, leading to gradient convergence/divergence issues. In the case of the BiLSTM model, which consists of multiple recurrent layers with forward and backward connections, the Kaiming Initialization function initializes the weights of both the forward and backward connections. It ensures that the weights of the recurrent connections are initialized in a way that allows information to flow effectively in both directions through the network during training. By initializing the weights appropriately, the Kaiming Initialization function helps in stabilizing the learning dynamics of the neural network. It facilitates smoother and more efficient training by preventing the gradients from becoming too small or too large, which can hinder the convergence of the optimization process. The Kaiming Initialization function combined with the BiLSTM algorithm can increase learning performance, accelerate convergence, and improve generalization to previously unknown material. It aids in overcoming the difficulties involved with training deep recurrent neural networks, particularly in tasks that require sequential data processing, such as natural language processing and time series analysis.

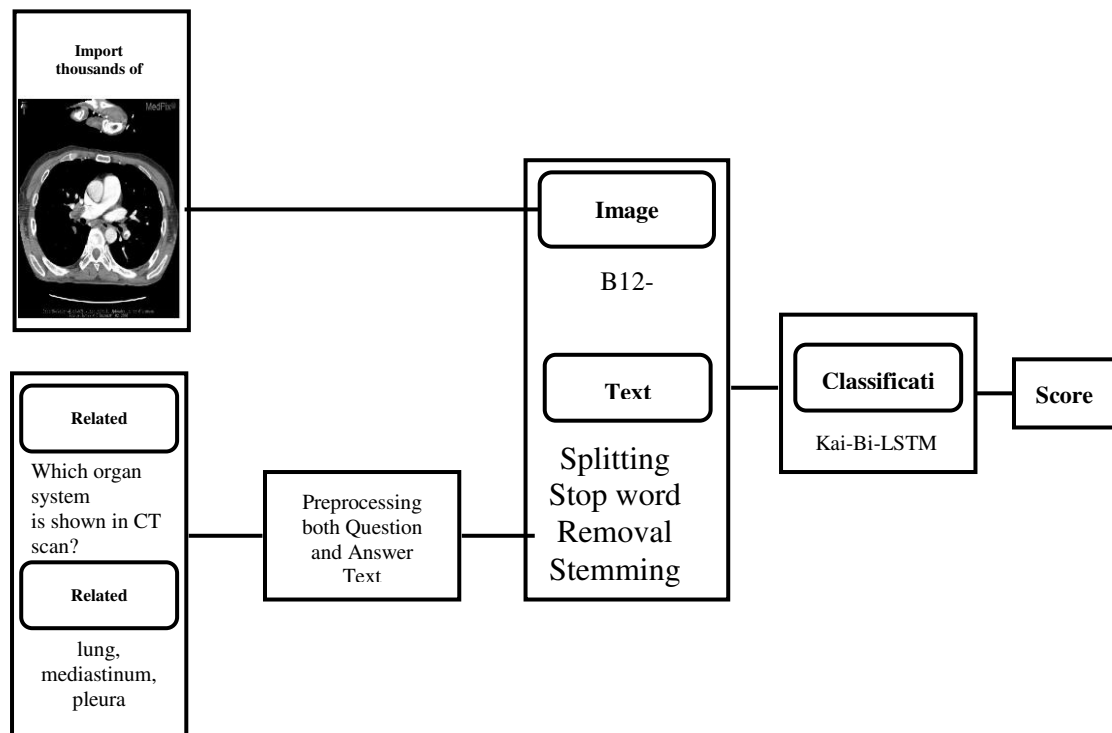


Fig 32 : Block diagram of QA System in the Training phase

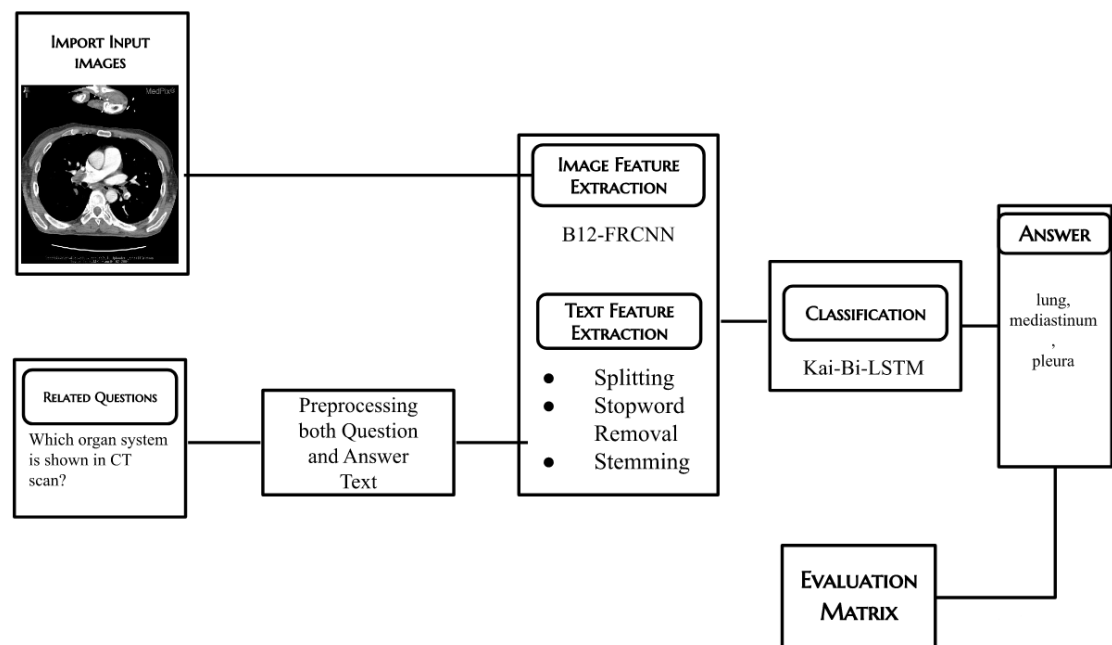


Fig 33 : Block diagram of QA System on Testing Phase

EXPERIMENTAL RESULT AND ANALYSIS



The experimental results showcase the accuracy of both the B12 FRCNN and Kai-BiLSTM models. Additionally, they highlight the comparative performance of the existing system, which combines skeletal image feature extraction via Convolutional Neural Network (CNN) and textual or question answer feature extraction using Long Short Term Memory (LSTM) models, exhibiting an accuracy of 83.9 percent across visual and textual datasets.

Existing feature detection algorithms, leveraging deep belief networks (DBN) and LSTM, demonstrate an accuracy rate of 85.9%. Various models employing LSTM for textual feature analysis and Recurrent Neural Networks (RNN) for feature extraction achieve an accuracy of 89.1946. Furthermore, the accuracy of other existing models, such as CNN and BiLSTM, analyzing visual and textual data, reaches 91.222%.

There is potential for further precision enhancement in the proposed model by leveraging updated datasets. Leveraging the new B12 FRCNN (Block12 Faster RCNN) model and the Kai-BiLSTM model, introduced with a remarkable 96.9% accuracy in this study, could significantly contribute to this improvement.

4.1 Experimental Result on VQA

The CLEF initiative labs are organizing the CLEF Image Retrieval and Classification Task 2019 campaign, inviting teams worldwide to participate in various research tasks. To ensure a focused evaluation, the questions are categorized based on modality, plane, organ system, and abnormality. These categories aim to challenge text creation and classification techniques effectively. Specifically, medical questions in this VQA challenge focus on individual characteristics, allowing for assessment solely based on visual content, without requiring specialized medical expertise.

The Healthcare Visual Q&A 2019 Training Set provides the most frequent answers for each category. For instance, modality options include xr-plain film, t2, us-ultrasound, and more. Similarly, the plane category lists axial, sagittal, coronal, and other planes. The organ system category comprises responses like skull and contents, musculoskeletal, gastrointestinal, and others. Finally, the abnormality category includes responses such as yes, no, meningioma, glioblastoma multiforme, and more.

To uphold precision, the responses generated during testing underwent manual validation by both a physician and a radiologist. A total of thirty-three responses were

adjusted, primarily to incorporate optional elements, enhance the range of viable responses, or refine automated replies. Because the training and validation sets were created using the same data generation procedures, the error rate should be similar. The test set includes 500 medical images and 500 related questions.

Evaluation metrics are crucial for assessing the performance, efficiency, and success of a system, process, or strategy. These metrics, which can be statistical or interpretive, serve as accurate measurements or indicators and are often based on key performance indicators (KPIs). They are widely utilized across various domains such as business, marketing, healthcare, education, and technology to evaluate the effectiveness of strategies or procedures, identify areas for improvement, and derive insights from collected data. In the realm of VQA research, a key goal is to develop computer vision systems capable of performing diverse tasks rather than specializing in just one area like object recognition.

The Precision metric in Visual Question Answering (VQA) quantifies the proportion of questions within a dataset for which the model generates correct answers. In VQA, the system receives an image along with a natural language question and is tasked with producing a suitable response.

In the context of Visual QA (VQA), the accuracy metric is often calculated by dividing the total number of right answers by the total number of questions in the dataset. For example, if a VQA model correctly answers 800 out of 1,000 questions in a dataset, its accuracy will be 80%. Accuracy is a crucial evaluation metric that indicates how well a model understands visual content and responds to related questions. To acquire a more complete knowledge of a model's performance, additional measures such as precision, recall, and F1 score are conceivably used. These measures provide nuanced insights beyond accuracy, allowing for a more complete assessment of the model's capabilities. The assessed accuracy of our model remains at approximately 50%.

By comparing generated translations to a reference translation, the metric known as BLEU (Bilingual Evaluation Understudy) is frequently used in machine translation to assess the standard of the output. To assess the effectiveness of visual question-answering (VQA) systems, BLEU has also been modified.

The BLEU score in VQA evaluates how closely the system's generated responses correspond to the reference responses given in the dataset. For each question-answer combination, the metric is first calculated by altering the n-gram precision, which quantifies the overlap of n-gram sequences between the generated and reference answers. The geometric mean of the n-gram precisions for each question in the dataset is then computed using the revised n-gram precision.

A higher BLEU (0-1) score indicates superior performance. Although the BLEU score can offer some indication of the quality of outputs from a VQA system, it is crucial to acknowledge its substantial limitations. These limitations include its failure to capture semantic similarity across responses and its inability to consider the diversity of valid answers to a given question. Consequently, it is advisable to employ a range of evaluation measures, including BLEU, to obtain a comprehensive understanding of the effectiveness of a VQA system.

$$\min(1 - \frac{r}{c}, 0) + \sum_{n=1}^4 \frac{\log p_n}{4} \quad (12)$$

where:

The reference_length attribute denotes the length of the reference answer, while the output_length attribute specifies the length of the generated answer. Blue_ngram_weights refer to the weights utilized for computing n-gram precisions, where p_n signifies the n-gram precision for n-grams of length n in the generated response. Here, n represents the length of the n-grams employed to determine precision.

The weights utilized to compute n-gram precisions are typically predetermined, although they can alternatively be derived from data. For instance, if unigrams ($n=1$) and bigrams ($n=2$) are selected, the weights may be assigned as [0.5, 0.5] to evenly distribute the significance of each precision.

The geometric mean of the BLEU score is calculated by aggregating the corrected n-gram precisions obtained from each item in the dataset. The discrepancy in length between the reference and generated responses is factored in during the computation of the adjusted n-gram precision. Specifically, it is computed as the exponentiation of the arithmetic mean of the log-transformed n-gram precisions. The outcome analysis

of the first phase dataset Healthcare Visual Q&A 2019 for each type of Visual Question Answering System.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Fig 34 : Confusion Matrix

The F-measure, often known as the F-score or F1 metric when the β value is 1, is a weighted harmonic mean of Recall and Precision. This metric is utilized for several reasons. The harmonic mean is typically the appropriate choice when averaging rates or frequencies. Additionally, a set-theoretic rationale for its use will be addressed subsequently. The more general form denoted as F allows for variable weighting of Recall and Precision, although it is common practice to assign them equal weight, resulting in the F1 score, which is the prevalent reference when discussing the F-measure.

A variant of accuracy not affected by negatives, single value measures(compare, tune systems). Harmonic mean of P and R is mentioned in equation (12)

$$F_{\beta} = \frac{(\beta^2+1).P.R}{\beta^2 P+R} \quad (13)$$

where $\beta = 1$, which gives

$$F_1 = \frac{2PR}{P+R} \quad (14)$$

Geometric interpretation the percentage overlap between relevant and retrieved which followed by

$$F_1 = \frac{2PR}{P+R} = 2\left(\frac{1}{P} + \frac{1}{R}\right)^{-1} \quad (15)$$

$$F_1 = 2\left(\frac{TP+FP}{TP} + \frac{TP+FN}{TP}\right)^{-1} = 2\frac{rel.ret}{rel+ret} \quad (16)$$

Precision is a statistical metric utilized to assess the accuracy of positive predictions generated by a classifier. It is calculated as the quotient of true positive predictions divided by the total number of positive predictions made by the classifier, irrespective of their correctness.

The formula for precision is:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Another way to describe accuracy is as the percentage of correctly predicted positive cases (true positives) out of all instances identified as positive by the classifier.

Here's a breakdown of the terms used in the formula:

- True Positives (TP): The number of occurrences accurately recognised as positive by the classifier that are also true positives.
- False Positives (FP): The number of cases that the classifier wrongly classified as positive despite being negative.

To calculate precision, you need to count the number of true positives and false positives from the classifier's predictions and then plug them into the formula. The calculated ratio will fall within the range of 0 to 1, where higher values correspond to higher precision, indicating fewer false positive predictions made by the classifier.

Recall, also known as sensitivity or true positive rate, is a statistic that assesses a classifier's ability to properly identify positive occurrences among all actual positive examples in a dataset. It calculates the proportion of true positive predictions made by the classifier compared to the total number of actual positive cases.

The formula for recall is:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Recall is defined as the ratio of accurately predicted positive cases (true positives) to total positive instances (true positives + false negatives).

Here's a breakdown of the terms used in the formula:

- True Positives (TP): This refers to the count of instances that are genuinely positive and are accurately identified as such by the classifier.
- False Negatives (FN): This indicates the count of positive occurrences that the classifier incorrectly identifies as negative.

To calculate recall, you need to count the number of true positives and false negatives from the classifier's predictions and then plug them into the formula. The resulting value will be a ratio between 0 and 1, where a higher value indicates better recall (i.e., fewer false negatives).

The F1 score, often known as the F-measure or F-score, is a metric for assessing the effectiveness of a classification model. It takes into account both the model's precision and recall in order to compute a single score that balances the trade-offs.

The F1 score is calculated using the following formula:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Where:

- Precision is the percentage of accurate positive predictions from all positive predictions provided by the model. It assesses the accuracy of optimistic predictions.
- Recall is the proportion of true positive predictions among all positive instances in the dataset. It evaluates the model's ability to correctly identify positive cases.
- The F1 score represents the harmonic mean of precision and recall. It ranges from 0 to 1, with higher values indicating better performance.

In classification, sensitivity, also referred to as recall or true positive rate, evaluates the classifier's capability to correctly identify positive instances from the entirety of actual positive examples within the dataset. Sensitivity holds particular importance in scenarios where the cost associated with overlooking positive examples (false negatives) is significant. It serves as a metric to gauge the classifier's effectiveness in capturing all pertinent instances of a specified class.

Mathematically, sensitivity is calculated using the following formula:

$$\text{Sensitivity (Recall)} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

- True Positives (TP) are instances that the model effectively classifies as positive.
- False Negatives (FN) are situations that are genuinely positive but are misclassified as negative by the model.

Sensitivity is a fraction that runs between 0 and 1, with a number closer to 1 indicating more sensitivity or memory. A sensitivity of one suggests that the classifier properly recognises all positive examples, whereas a sensitivity of zero indicates that the classifier fails to identify any positive instances.

Sensitivity is widely utilized in medical diagnostics, anomaly detection, and other applications where identifying true positives is critical.

Classification specificity evaluates a classifier's capacity to accurately recognize negative instances among all genuine negative examples in a dataset. It serves as a complement to the false positive rate and is particularly advantageous in scenarios where the consequences of false alarms (false positives) are significant.

Mathematically, specificity is calculated using the following formula:

$$\text{Classification Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

Where:

- True Negatives (TN) are events that were accurately categorized as negative by the model.
- False Positives (FP) are events that are truly negative but are misclassified as positive by the model.

Specificity, represented as a fraction between 0 and 1, signifies the degree of accuracy in identifying negative instances by the classifier. A value closer to 1 suggests higher specificity, indicating that the classifier adeptly detects all negative occurrences. Conversely, a specificity value nearing zero implies the classifier fails to identify any negative examples. This metric holds significance in various domains, including

medical diagnostics and spam detection, where minimizing false alarms is paramount for effective decision-making.

Table 6 : Most Frequent Answers for various Question Type and Answer count

Question Types	Most Frequent Answers	Total No. of Answers
MODALITY	No	554
	Yes	552
	xr-plain film	456
	t2	217
	us-ultrasound	183
	t1	137
	contrast	107
	noncontrast	102
	ct non contrast	84
PLANES	axil	1558
	sagittal	478
	coronal	389
	ap	197
	lateral	151
	frontal	120
	pa	92
	transverse	76
	oblique	50
ORGAN SYSTEM	genitourinary	214
	face, sinuses and neck	191
	vascular and lymphatic	122
	heart and great vessels	120
	breast	65
	musculoskeletal	438
	Yes	62
	No	48

Question Types	Most Frequent Answers	Total No. of Answers
ABNORMALITY	meningioma	30
	glioblastoma multiforme	28
	pulmonary embolism	16
	acute appendicitis	14
	arteriovenous malformation (avm)	14
	arachnoid cyst	13
	schwannoma	13
	tuberous sclerosis	12
	brain, cerebral abscess	12
	ependymoma	12
	fibrous dysplasia	12
	multiple sclerosis	12
	diverticulitis	11
	langerhan cell histiocytosis	11
	sarcoidosis	11

Total No. of Answers vs. Most Frequent Answers

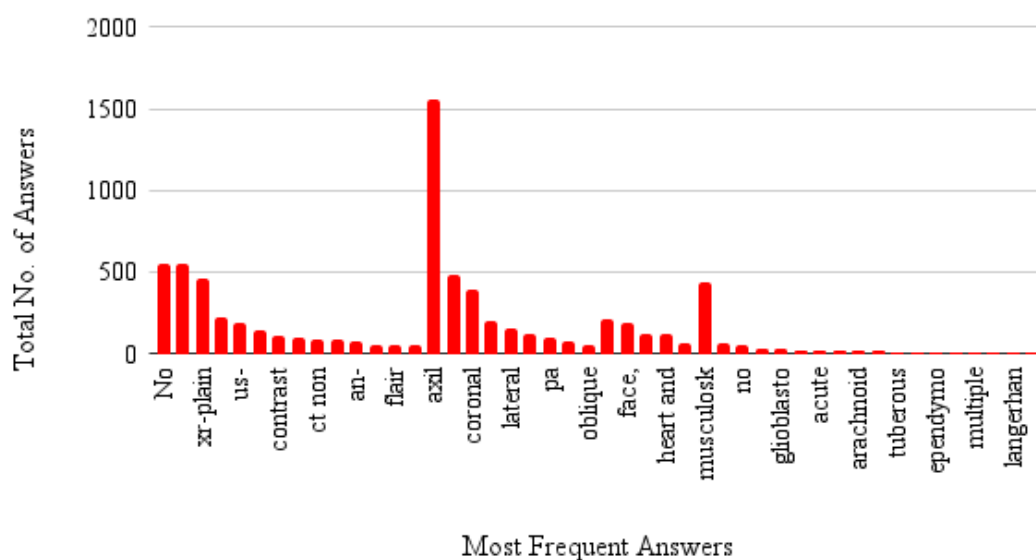


Fig 33 : Most frequent answers from different question types

Modality, Plane, Organ and Abnormality

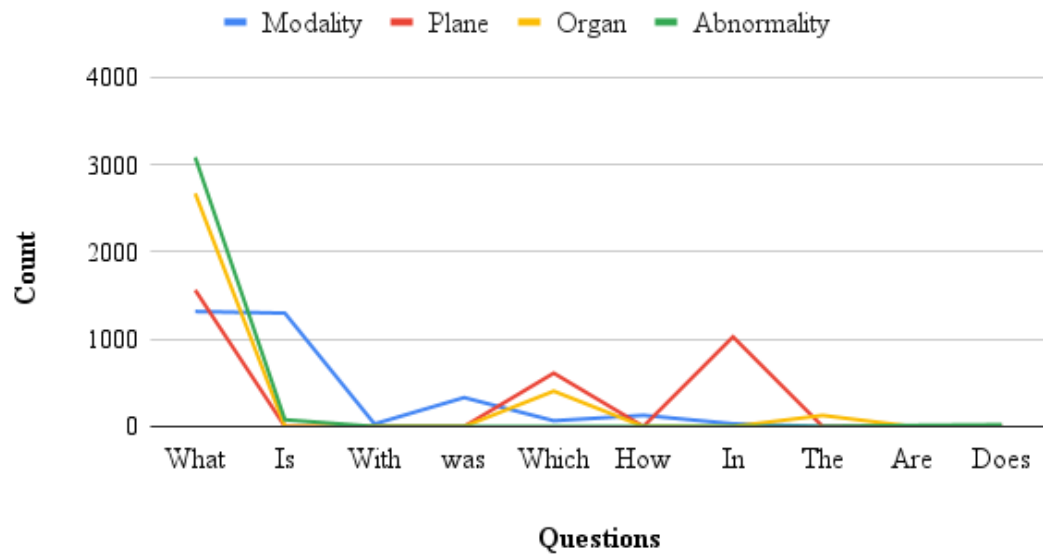


Fig 36 : Different methods of question for various question types

Table 7 : Total count of Visual and Textual dataset

Images	3200
No. of Questions and Answer	12792

Visual and Textual Dataset Count

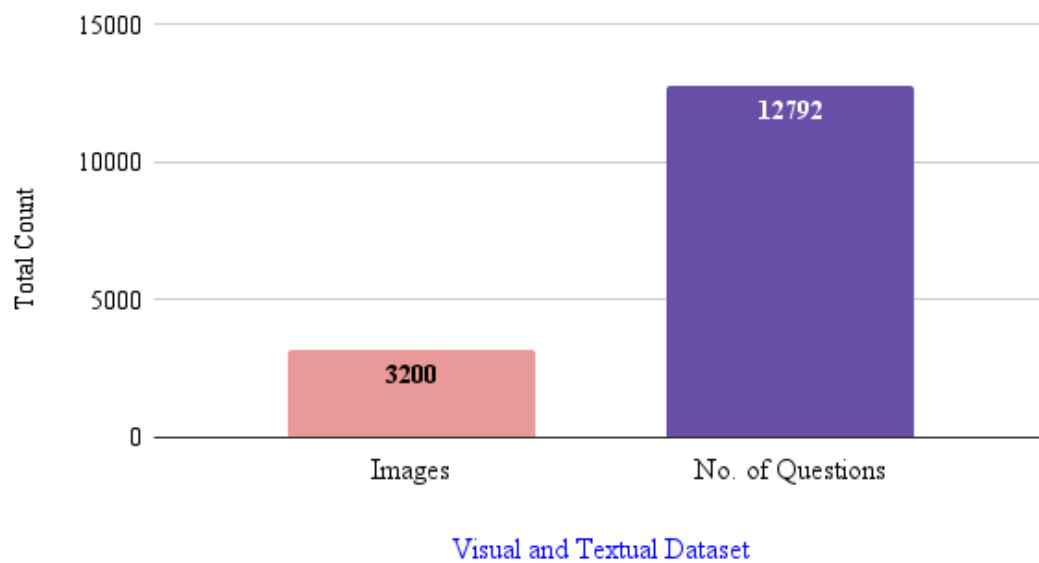
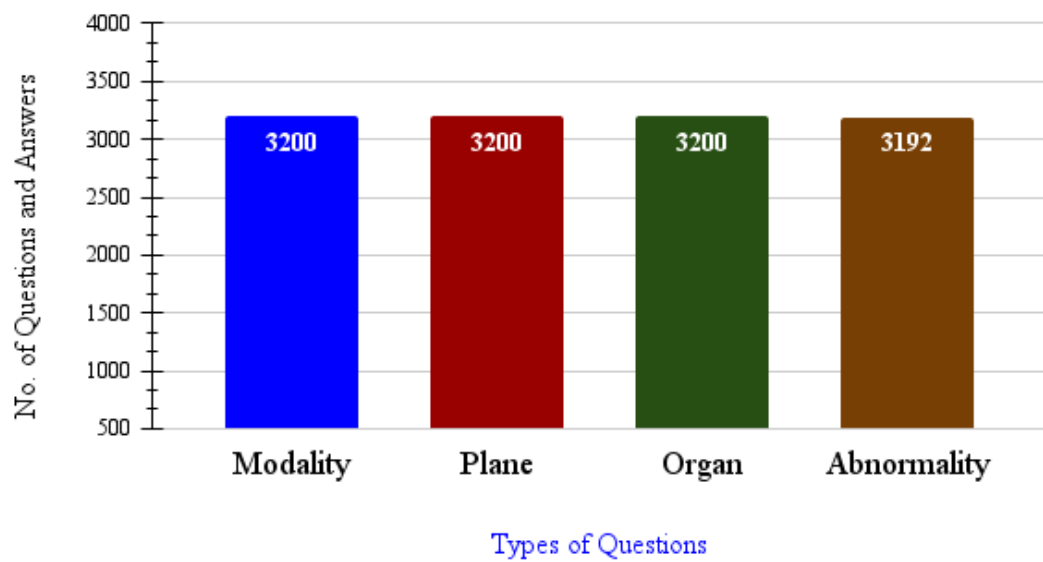
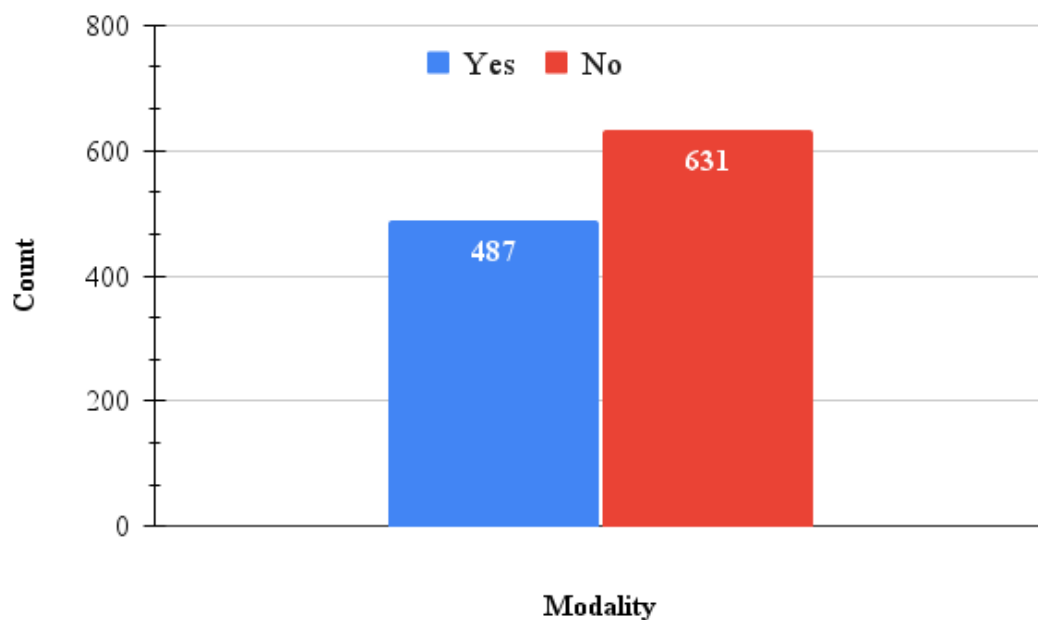


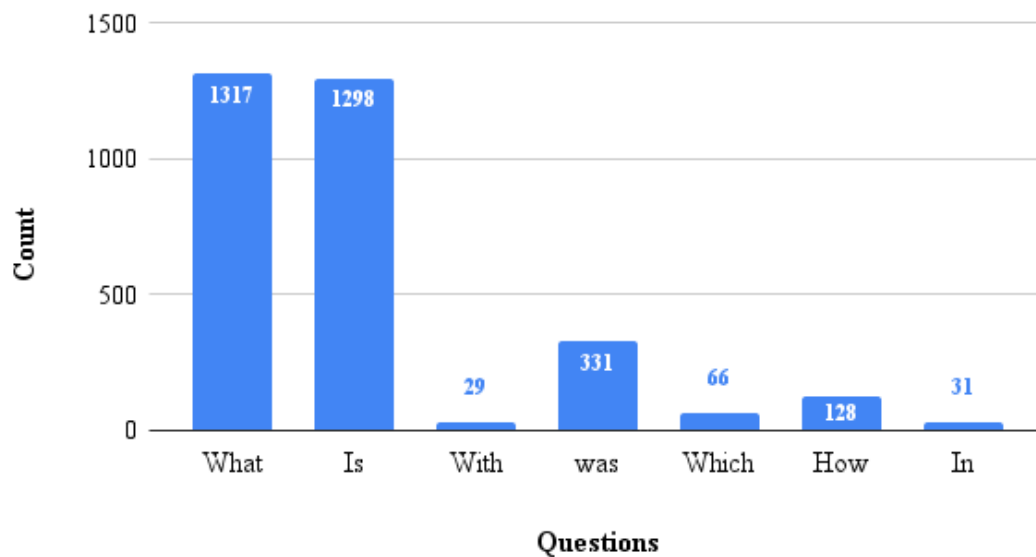
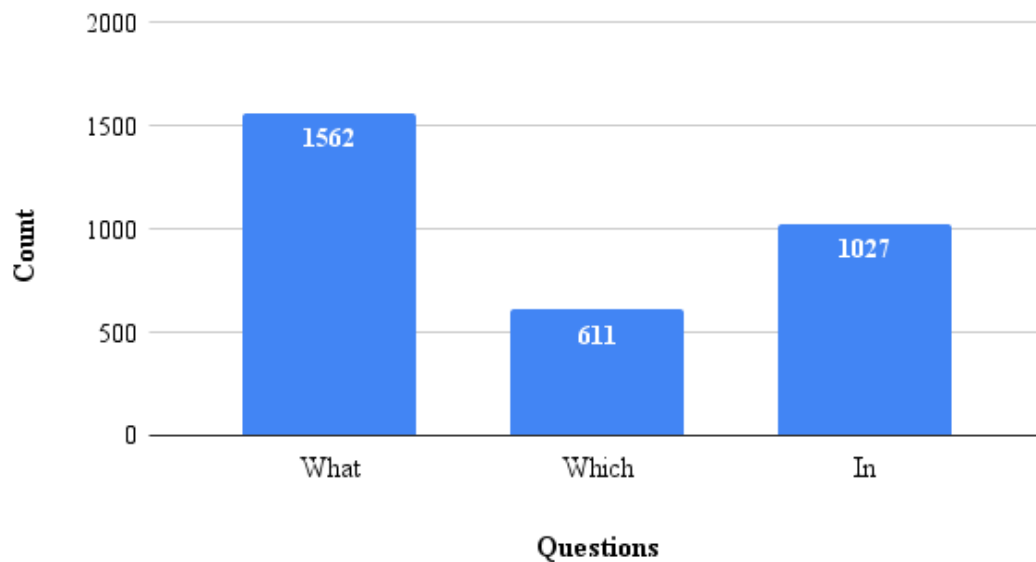
Fig 37 : Total count of Image and questions from CLEF Image Retrieval and Classification Task 2019 Dataset

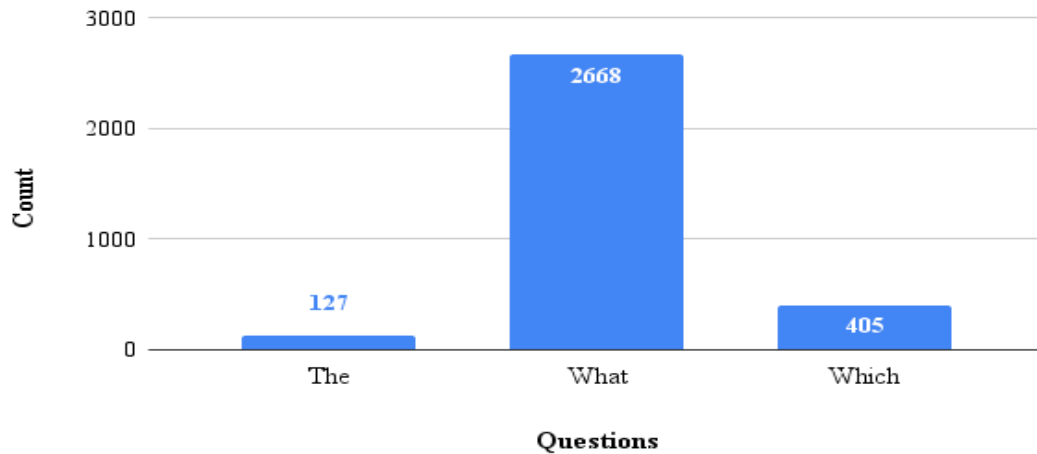
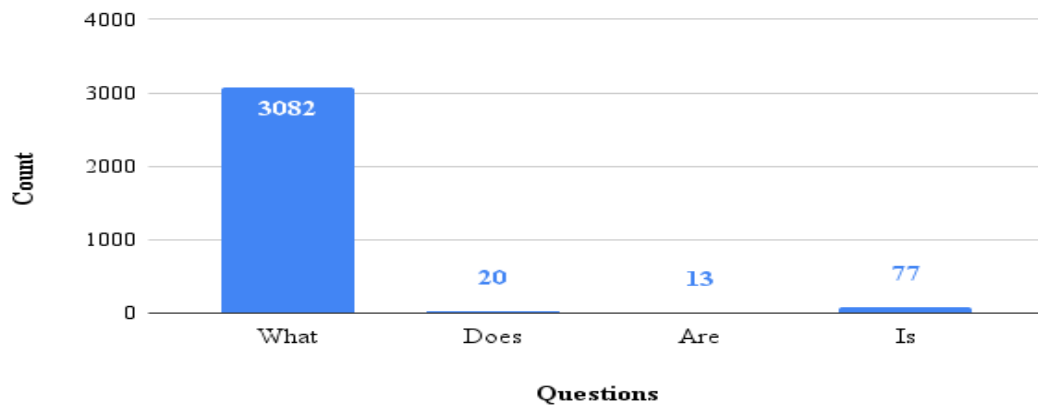
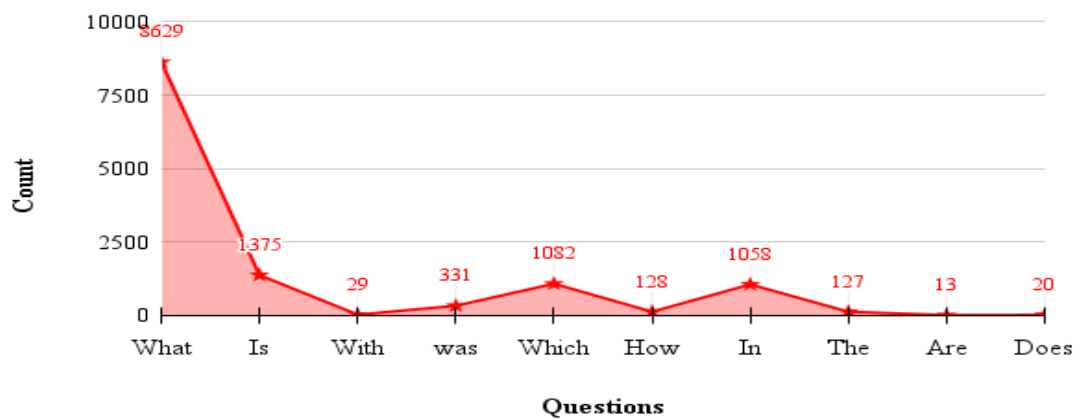
The table provided indicates the distribution of questions and answers across different types of questions, namely Modality, Plane, Organ, and Abnormality. Here's a detailed explanation:

1. **Modality:** This category pertains to questions related to the imaging modality used to capture medical images, such as X-ray, MRI, CT scan, ultrasound, etc. The table indicates that there are a total of 3200 questions and answers associated with the Modality category.
2. **Plane:** Refers to questions concerning the imaging plane or orientation of the captured medical images, such as axial, sagittal, coronal, etc. Similar to the Modality category, there are also 3200 questions and answers related to the Plane category.
3. **Organ:** Represents questions regarding specific organs or organ systems depicted in the medical images. Examples include questions about the brain, lungs, heart, musculoskeletal system, etc. Again, there are 3200 questions and answers allocated to the Organ category.
4. **Abnormality:** Denotes questions pertaining to the identification or diagnosis of abnormalities or pathologies present in the medical images. This could include conditions like tumors, fractures, infections, etc. There are slightly fewer questions and answers in this category, totaling 3192.

Overall, each category has an equal number of questions and answers, except for the Abnormality category, which has a slightly lower count. These questions and answers are crucial for training and evaluating models in medical image analysis and interpretation tasks, contributing to advancements in healthcare and diagnostic capabilities.

No. of Questions and Answers vs. Types of Questions**Fig 38 : Total number of data from both question and answer for various types****Fig 39 : Count of boolean questions for Modality**

Questions and its count for Modality**Fig 40 : Various types of question for Modality question answering data type****Question and its count for plane****Fig 41 : Various types of question for Plane question answering data type**

Question and its count for Organ**Fig 42 : Various types of question for Organ question answering data type****Question and its count for Abnormalites****Fig 43 : Various types of question for Abnormality question answering data type****Total Count****Fig 44 : Overall count of different question types**

<pre>supported_activation_functions = ("sigmoid", "relu", "softmax") def sigmoid(self, sop): if type(sop) in [list, tuple]: sop = numpy.array(sop) return 1.0 / (1 + numpy.exp(-1 * sop)) def relu(self, sop): if not (type(sop) in [list, tuple, numpy.ndarray]): if sop < 0: return 0 else: return sop elif type(sop) in [list, tuple]: sop = numpy.array(sop) result = sop result[sop < 0] = 0 return result def softmax(self, layer_outputs): return layer_outputs / (numpy.sum(layer_outputs) + 0.000001)</pre>	<pre>def layers_weights(self, model, initial=True): network_weights = [] layer = model.last_layer while "previous_layer" in layer.__init__.__code__.co_varnames: if type(layer) in [self.Conv2D, self.Dense]: if initial == True: network_weights.append(layer.initial_weights) elif initial == False: network_weights.append(layer.trained_weights) else: raise ValueError("Unexpected value to the 'initial' parameter: initial").format(initial=initial)) # Go to the previous layer. layer = layer.previous_layer # If the first layer in the network is not an input layer (i.e. an instance of the Input2D class), raise an error. if not (type(layer) is self.Input2D): raise TypeError("The first layer in the network architecture must be an input layer.") network_weights.reverse() return numpy.array(network_weights)</pre>
<pre>weights_vector = vector_weights[start:start + layer_weights_size] # matrix = pygad.nn.DenseLayer.to_array(vector=weights_vector, shape=layer_weights_shape) matrix = numpy.reshape(weights_vector, newshape=(layer_weights_shape)) network_weights.append(matrix) start = start + layer_weights_size # Go to the previous layer. layer = layer.previous_layer # If the first layer in the network is not an input layer (i.e. an instance of the Input2D class), raise an error. if not (type(layer) is self.Input2D): raise TypeError("The first layer in the network architecture must be an input layer.") network_weights.reverse() return numpy.array(network_weights) def layers_weights_as_vector(self, model, initial=True): network_weights = [] layer = model.last_layer while "previous_layer" in layer.__init__.__code__.co_varnames: if type(layer) in [self.Conv2D, self.Dense]: # If the 'initial' parameter is True, append the initial weights. Otherwise, append the trained weights.</pre>	<pre>if not (type(layer) is self.Input2D): raise TypeError("The first layer in the network architecture must be an input layer") network_weights.reverse() return numpy.array(network_weights) def update_layers_trained_weights(self, model, final_weights): layer = model.last_layer layer_idx = len(final_weights) - 1 while "previous_layer" in layer.__init__.__code__.co_varnames: if type(layer) in [self.Conv2D, self.Dense]: layer.trained_weights = final_weights[layer_idx] layer_idx = layer_idx - 1 # Go to the previous layer. layer = layer.previous_layer</pre>

Fig 45 : Sample code function on Existing CNN algorithm

4.2 Code Implementation

This Python class **ExistingCNN** contains several methods related to a convolutional neural network (CNN) model, such as activation functions (**sigmoid**, **relu**, **softmax**), weight manipulation, and updating trained weights. Here's a brief explanation of each function:

1. **Activation Functions:** The class provides implementations for common activation functions used in neural networks, including sigmoid, ReLU, and softmax.
2. **layers_weights:** This method extracts the weights of all layers in the model and returns them as an array. It can return either the initial weights or the trained weights depending on the **initial** parameter.

3. `layers_weights_as_matrix`: Similar to `layers_weights`, but it returns the weights of each layer reshaped as a matrix instead of an array.
4. `layers_weights_as_vector`: Similar to `layers_weights`, but it returns the weights of each layer flattened into a vector.
5. `update_layers_trained_weights`: This method updates the trained weights of each layer in the model using the provided `final_weights` array.

```
# construct DBN
dbn = ExistingDBN(input=x, label=y, n_ins=6, hidden_layer_sizes=[3, 3], n_outs=2, rng=rng)

# pre-training (TrainUnsupervisedDBN)
dbn.pretrain(lr=pretrain_lr, k=1, epochs=pretraining_epochs)

# fine-tuning (DBNSupervisedFineTuning)
dbn.finetune(lr=finetune_lr, epochs=finetune_epochs)

# test
x = numpy.array([[1, 1, 0, 0, 0, 0],
                 [0, 0, 0, 1, 1, 0],
                 [1, 1, 1, 1, 1, 0]])

# print
dbn.predict(x)

def training(self, iptrdata):
    parser = argparse.ArgumentParser(description='Train')

    parser.add_argument('-train', help='Train data', type=str, required=True)
    parser.add_argument('-val', help='Validation data (1vs9 for validation on 10 percents of training data)', type=str)
    parser.add_argument('-test', help='Test data', type=str)

    parser.add_argument('-e', help='Number of epochs', type=int, default=1000)
    parser.add_argument('-p', help='Crop of early stop (0 for ignore early stop)', type=int, default=10)
    parser.add_argument('-b', help='Batch size', type=int, default=128)

    parser.add_argument('-pre', help='Pre-trained weight', type=str)
    parser.add_argument('-name', help='Saved model name', type=str, required=True)

    train_inputs = []
    train_outputs = []
    time.sleep(78)
    if len(train_inputs) > 0:
        if (train_inputs.ndim != 4):
            raise ValueError("The training data input has {num_dims} but it must have 4 dimensions. The first dimension is the number of t
        if (train_inputs.shape[0] != len(train_outputs)):
            raise ValueError(
                "Mismatch between the number of input samples and number of labels: {num_samples_inputs} != {num_samples_outputs}."

    network_predictions = []
    network_error = 0
    for epoch in range(self.epochs):
        print("Epoch {epoch}".format(epoch=epoch))
        for sample_idx in range(train_inputs.shape[0]):
            # print("Sample {sample_idx}".format(sample_idx=sample_idx))
            self.feed_sample(train_inputs[sample_idx, :])

            try:
                predicted_label = \
                    self.numpy.where(self.numpy.max(self.last_layer.layer_output) == self.last_layer.layer_output)[0][0]
            except IndexError:
                print(self.last_layer.layer_output)
                raise IndexError("Index out of range")
            network_predictions.append(predicted_label)

            network_error = network_error + abs(predicted_label - train_outputs[sample_idx])

        self.update_weights(network_error)
```

Fig 46 : Construction of Existing DBN structure for Medical Images

1. **DBN Initialization:** An illustration of the `ExistingDBN` class is delivered with specified parameters such as input size (`n_ins`), number of hidden layers and their sizes (`hidden_layer_sizes`), output size (`n_outs`), and random number generator (`rng`).
2. **Pre-training:** The DBN is pre-trained using unsupervised learning (contrastive divergence algorithm) to learn the weights in an unsupervised manner. The learning rate (`pretrain_lr`) and number of pre-training epochs (`pretraining_epochs`) are specified.
3. **Fine-tuning:** After pre-training, the DBN is fine-tuned using supervised learning (backpropagation) to adjust the weights based on labeled data. The learning rate (`finetune_lr`) and number of fine-tuning epochs (`finetune_epochs`) are specified.
4. **Testing:** Test data (`x`) is provided to the trained DBN, and predictions are made using the `predict` method of the `ExistingDBN` class.
5. **Training Method:** The `training` method is defined, which appears to be used for training the DBN model. It parses command-line arguments for training data, validation data, test data, number of epochs, batch size, pre-trained weights, and saved model name.
6. **Data Validation:** The code checks if the training inputs have the correct dimensions and if the number of input samples matches the number of labels. If not, it raises a `ValueError`.
7. **Training Loop:** The code iterates over epochs and samples, feeds each sample to the network, makes predictions, calculates network error, and updates the weights based on the error using the `update_weights` method.

```

class Model:
    def __init__(self, last_layer, epochs=10, learning_rate=0.01):

        self.last_layer = last_layer
        self.epochs = epochs
        self.learning_rate = learning_rate

        # The network_layers attribute is a list holding references to all CNN layers.
        self.network_layers = self.get_layers()

    def get_layers(self):

        network_layers = []

        layer = self.last_layer

        while "previous_layer" in layer.__init__.__code__.co_varnames:
            network_layers.insert(0, layer)
            layer = layer.previous_layer

        return network_layers

    def train(self, train_inputs, train_outputs):

        if (train_inputs.ndim != 4):
            raise ValueError(
                "The training data input has {num_dims} but it must have 4 dimensions. The first dimension is the number of t
                num_dims=train_inputs.ndim))

        if (train_inputs.shape[0] != len(train_outputs)):
            raise ValueError(
                "Mismatch between the number of input samples and number of labels: {num_samples_inputs} != {num_samples_outp
                num_samples_inputs=train_inputs.shape[0], num_samples_outputs=len(train_outputs)))

        network_predictions = []
        network_error = 0

        for epoch in range(self.epochs):
            print("Epoch {epoch}".format(epoch=epoch))
            for sample_idx in range(train_inputs.shape[0]):
                # print("Sample {sample_idx}".format(sample_idx=sample_idx))
                self.feed_sample(train_inputs[sample_idx, :])

                try:
                    predicted_label = \
                        numpy.where(numpy.max(self.last_layer.layer_output) == self.last_layer.layer_output)[0][0]
                except IndexError:
                    print(self.last_layer.layer_output)
                    raise IndexError("Index out of range")
                network_predictions.append(predicted_label)

```

Fig 47 : Code function on Proposed Kai_BiLSTM

4.3 Comparison analysis of Feature extraction techniques

Table 8 : The table presents the performance metrics for a certain method across different evaluation criteria

Method	Accuracy
Manhattan Distance (MD)	50.40%
Euclidean Distance (ED)	57.80%
Jaccard Similarity Coefficient (JSC)	58.90%
Cosine Similarity (CS)	60.10%

This metric computes the ratio of correct predictions generated by the method. It quantifies the spatial separation between two points within a grid-based framework by summing the absolute disparities in their respective coordinates. Here, the method

achieved an accuracy of 50.40% when evaluated using Manhattan Distance. Euclidean Distance metric calculates the straight-line distance between two points in space. The method achieved an accuracy of 57.80% Jaccard Similarity Coefficient measures the similarity between two sets by comparing their intersection to their union. The method achieved an accuracy of 58.90% Cosine Similarity: This metric measures the angle between two vectors, indicating their similarity. The method achieved an accuracy of 60.10% as denoted in Table 8.

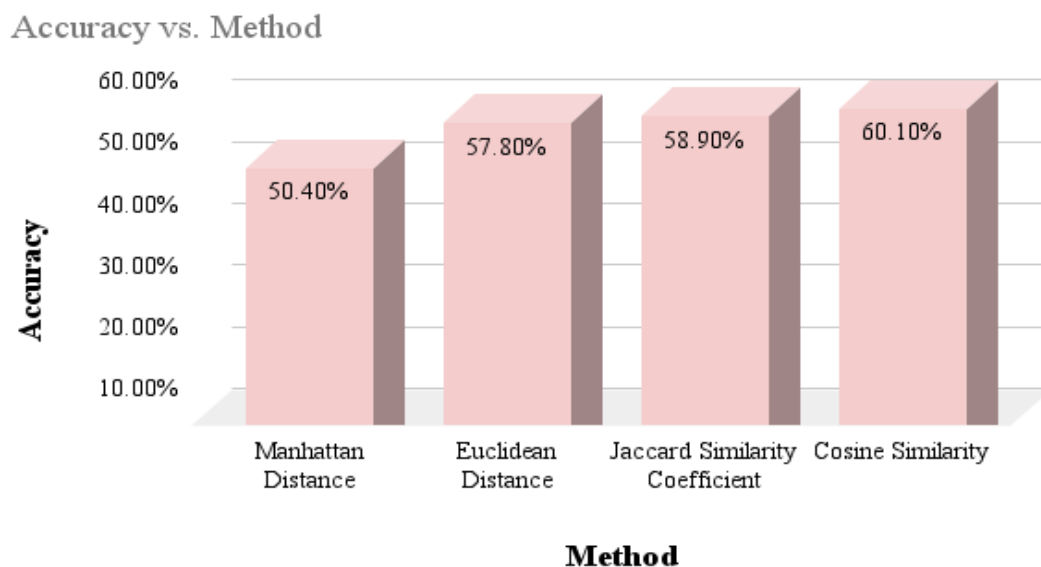


Fig 48 : The performance metrics for a certain method across different evaluation criteria

The KNN Classifier demonstrated an accuracy of 28.20 percent. KNN, or K-Nearest Neighbours, presents a straightforward approach to classifying data points by assigning them to the majority class among their nearest neighbors.

The Soft-Max Classifier achieved an accuracy of 34.10 percent. Widely employed in multiclass classification scenarios, the Soft-Max Classifier computes the probability distribution across all classes.

With an accuracy of 37.40 percent, the SVM Classifier employs Support Vector Machine techniques, particularly effective for binary and multiclass classification tasks, by determining optimal hyperplanes between classes.

The CNN Classifier, achieving the highest accuracy of 85.97 percent, relies on Convolutional Neural Network architecture tailored for image classification tasks. Through hierarchical feature extraction facilitated by convolutional layers, CNNs excel in discerning intricate patterns within images. These findings are summarized in Table 9.

Table 9 : The accuracy achieved by different classification algorithms

Classification Algorithm	Accuracy
K-NN Classifier	28.20%
Soft-Max Classifier	34.10%
SVM Classifier	37.40%
CNN Classifier	85.97%

Accuracy vs. Classification Algorithm

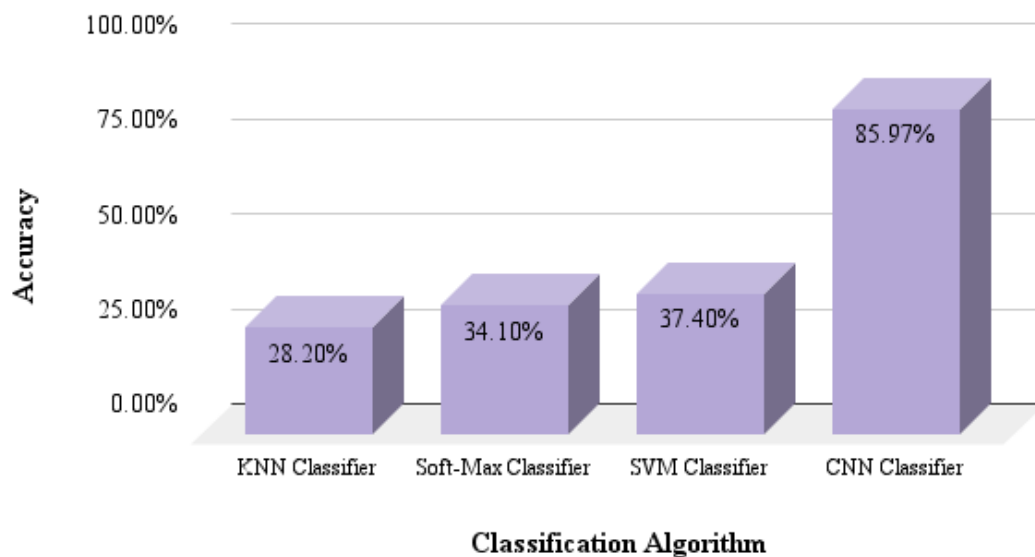


Fig 49 : The accuracy achieved by various classification techniques for current methodology

As mentioned in Table 10, The current BiLSTM achieved an F-measure of 91.23314. BiLSTM (Bidirectional Long Short-Term Memory) is a recurrent neural network architecture that is capable of capturing long-term dependencies in sequential data from both forward and backward directions. So for RNN it achieved an F-measure of 89.22156. RNN (Recurrent Neural Network) is a type of neural network architecture commonly used for sequential data processing tasks. For current DBN achieved an F-measure of 85.78629. The Deep Belief Network (DBN) is a generative neural network model composed of multiple layers of stochastic and latent variables. The Convolutional Neural Network (CNN) achieved an F-measure of 84.09091. CNN, short for Convolutional Neural Network, is a sophisticated deep learning architecture specifically designed for processing structured grid data, such as images.

Table 10 : The table shows the F-measure achieved by different algorithms

Algorithms	FMeasure
Existing BiLSTM	91.23314
Existing RNN	89.22156
Existing DBN	85.78629
Existing CNN	84.09091

Conjunction with various feature extraction approaches

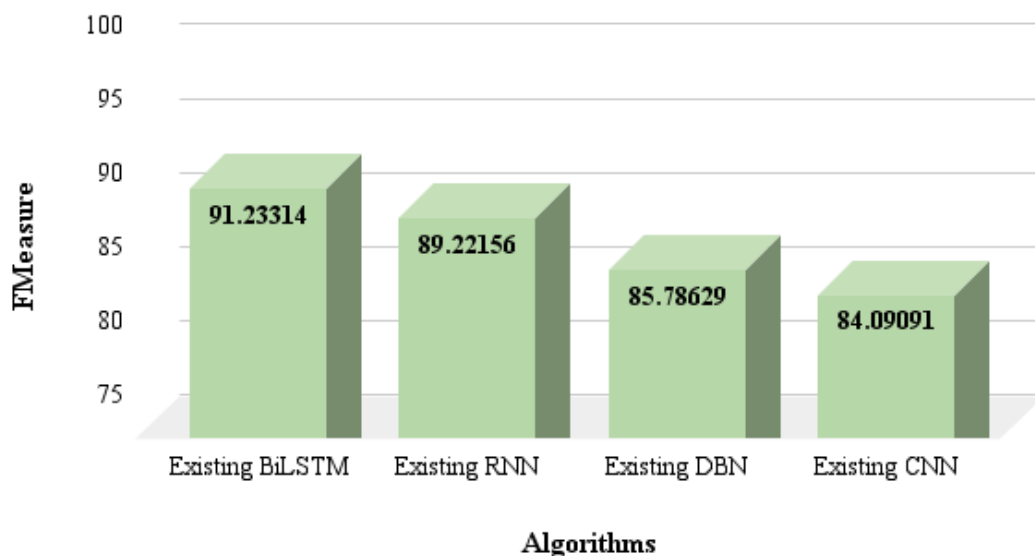


Fig 50 : F-measure achieved by different algorithms

Table 11 highlights the comparison of different algorithms which exist and with proposed algorithms. The convolutional neural network (CNN), dense based network (DBN), recurrent neural network (RNN) and Bidirectional LSTM are the current algorithms which are used for analysis. In which the highest F Measure is for the BiLSTM model. But when it is compared with the proposed Bidirectional LSTM the accuracy is high which is at 96.9%. So the proposed system predicts accurate output. This leads to the diagnostic system being very high quality.

Table 11 : The table compares the accuracy of both Proposed and Current algorithms

Algorithm	Accuracy
Proposed BiLSTM	96.9
Existing BiLSTM	91.33333
Existing RNN	89.1946
Existing DBN	85.9
Existing CNN	83.9

Accuracy vs. Algorithm

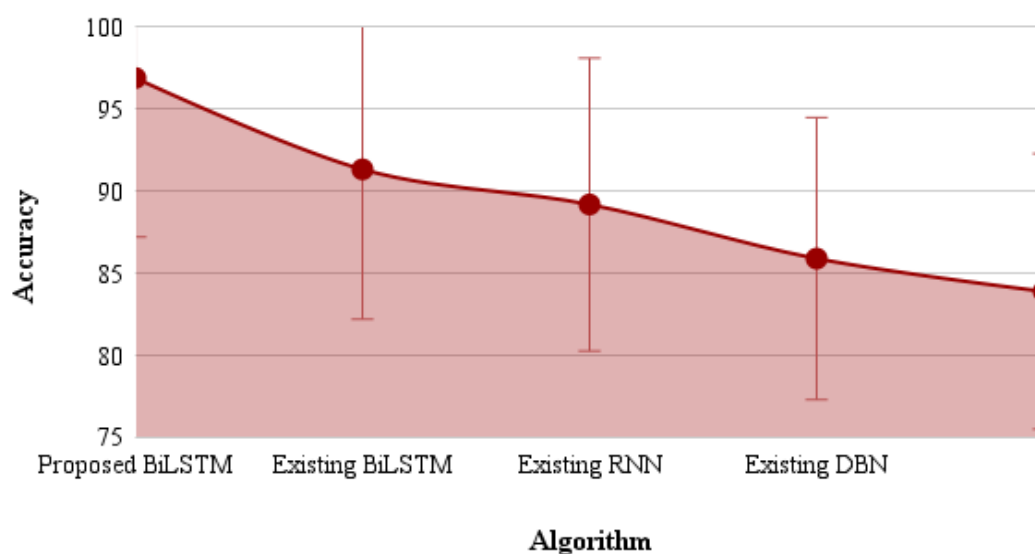


Fig 51 : Comparison with existing algorithm and proposed algorithm

Table 12 : The comparative analysis between Existing CNN and Existing BiLSTM with Proposed BiLSTM

Algorithm	Accuracy
Proposed BiLSTM	96.9
Existing BiLSTM	91.23
Existing CNN	85.97

Comparison Result of Proposed and Existing Algorithms

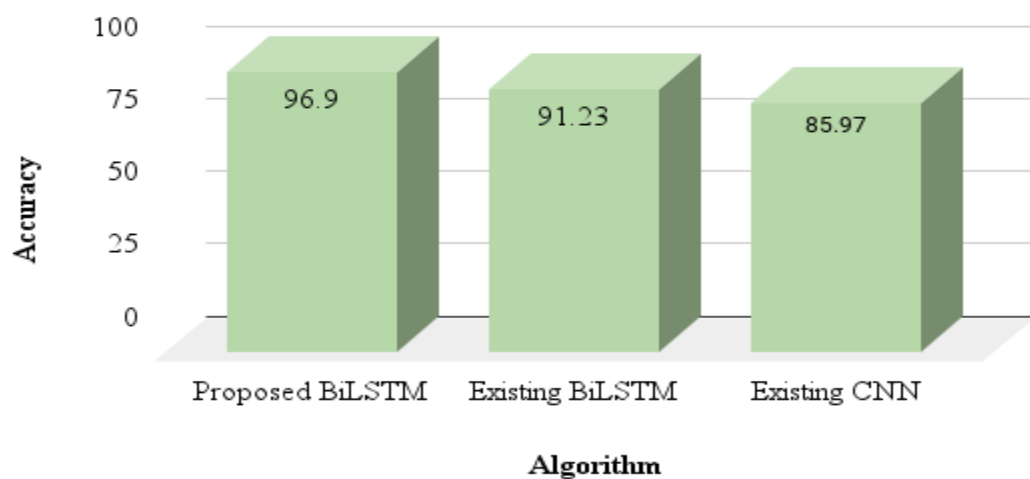


Fig 52 : The Comparison Result of Proposed and Existing algorithms

4.4 Snapshot on demonstration

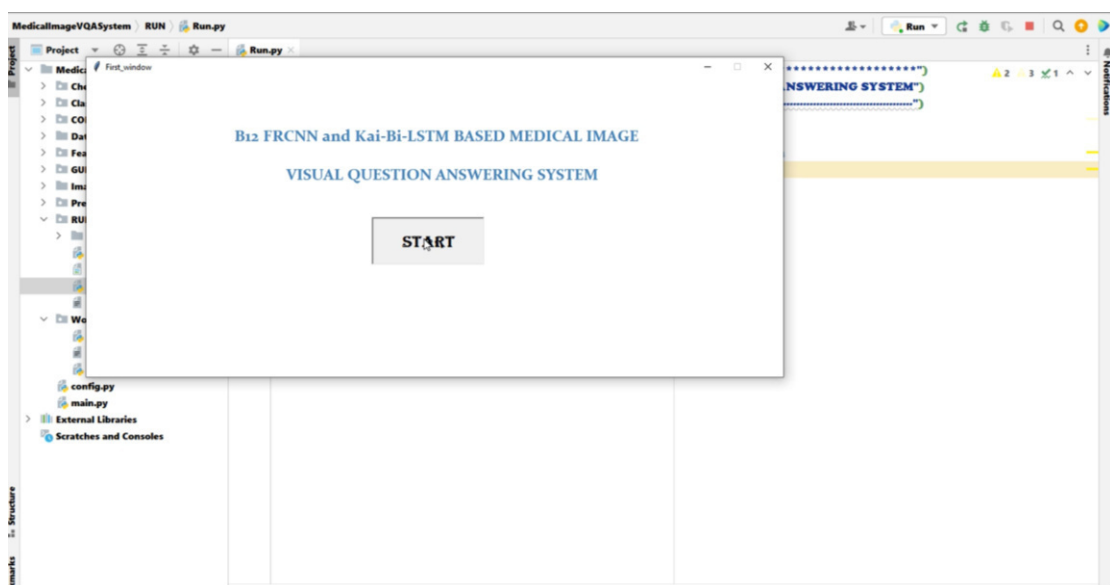


Fig 53 : Starting page of Application

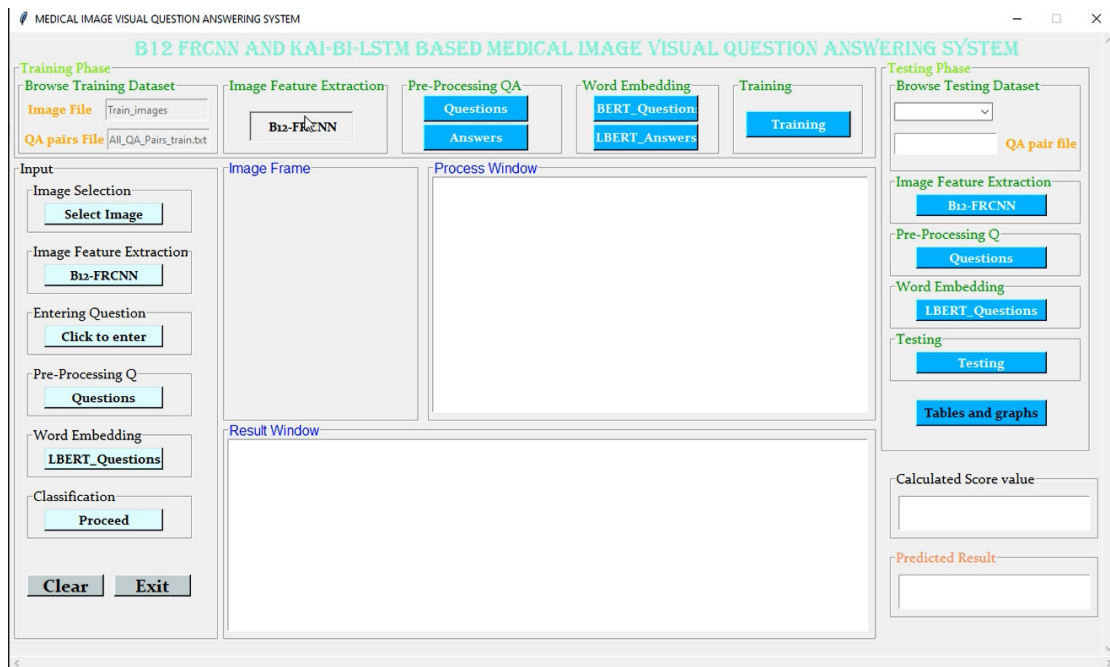


Fig 54 : Load the Training dataset both Image and Question Answer pair

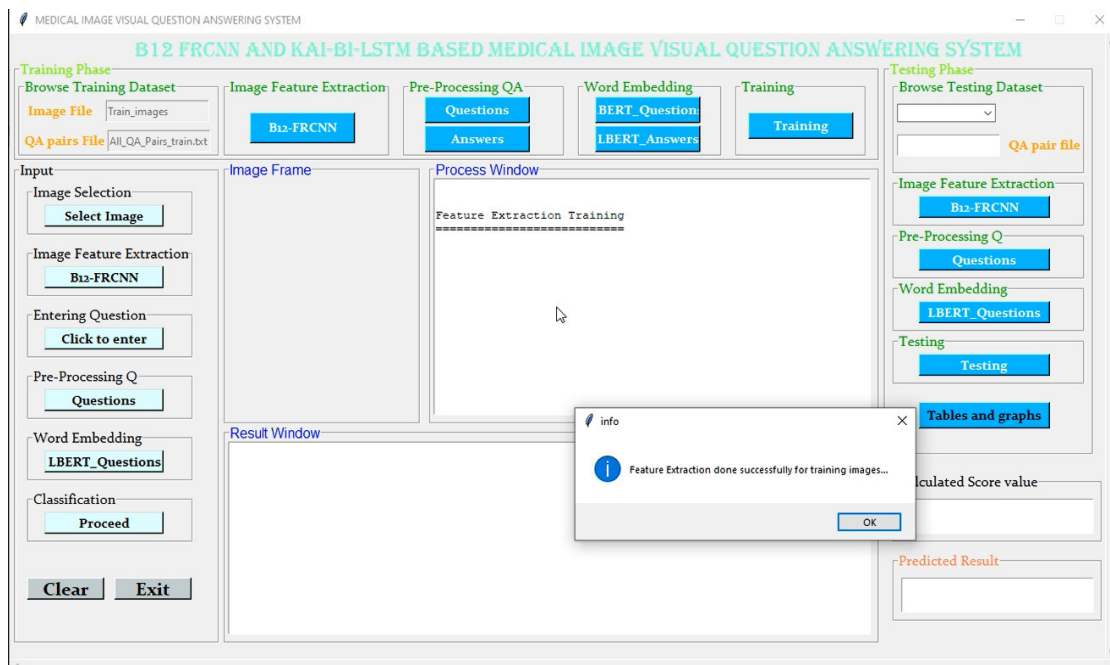


Fig 55 : Feature Extraction for Image dataset

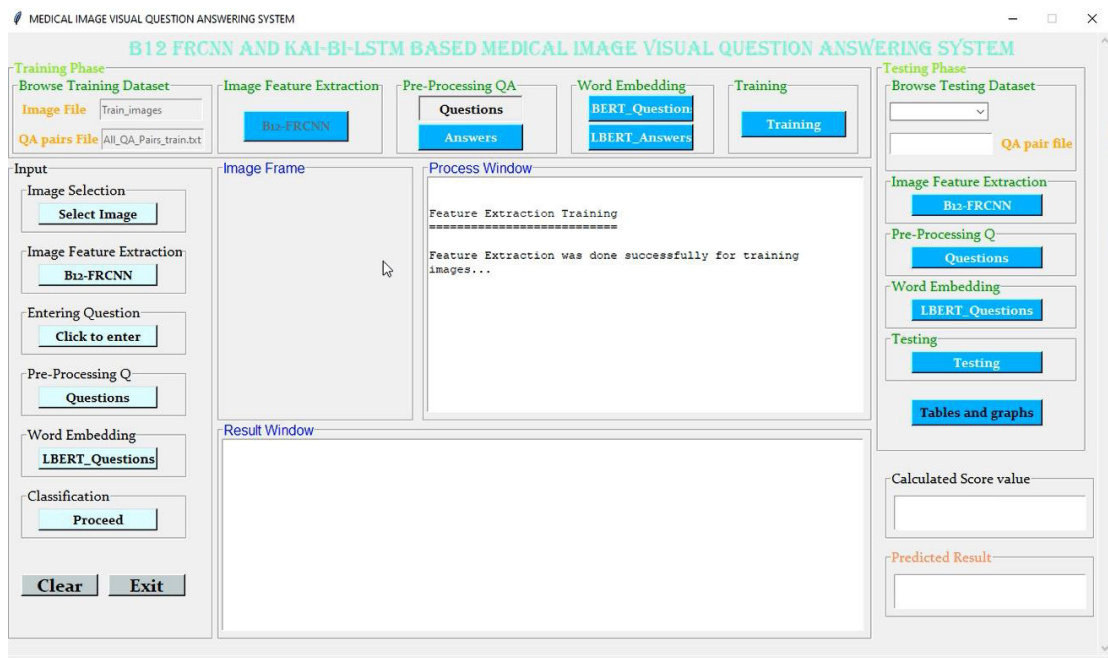


Fig 56 : Pre-process the Question dataset

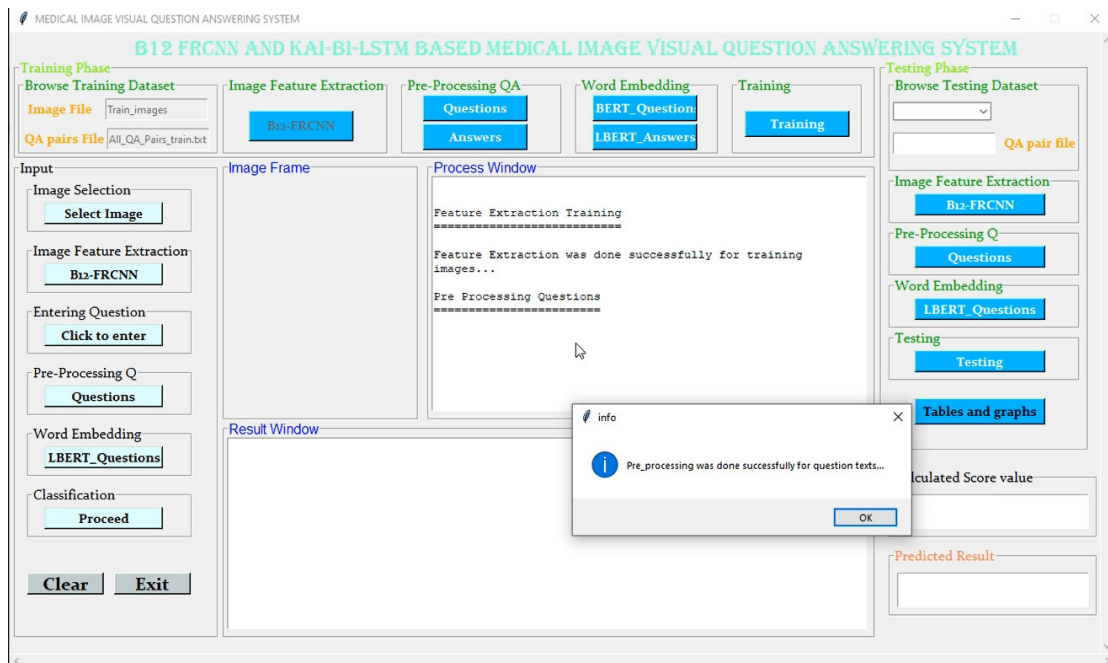


Fig 57 : Pre-process for Answer Dataset

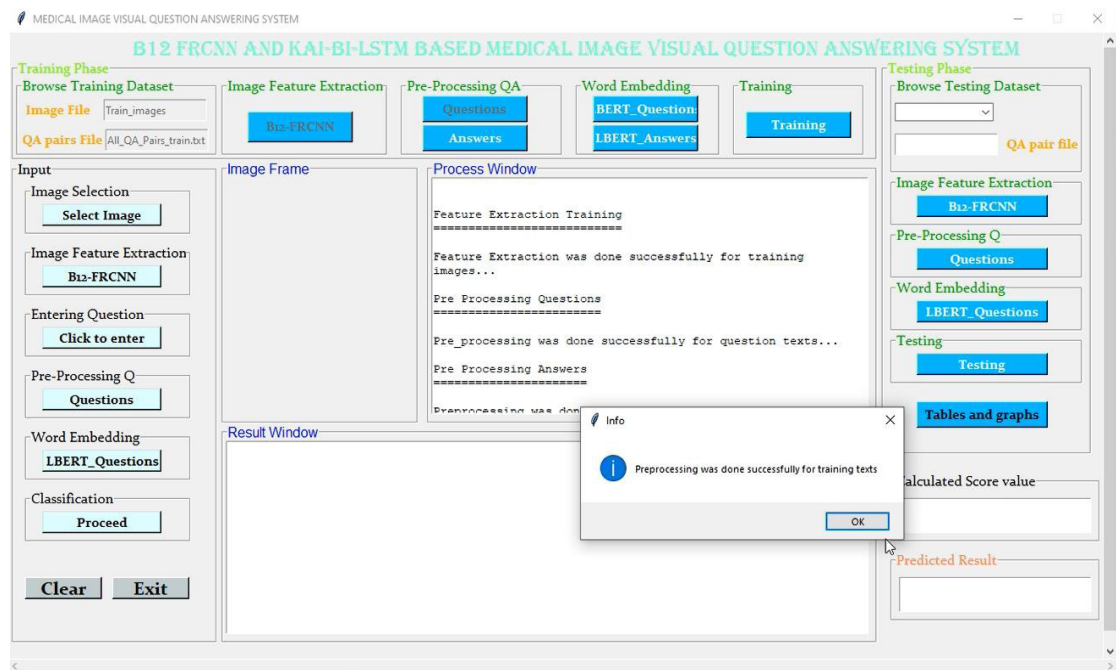


Fig 58 : Done with Preprocessing for both Question and Answer dataset

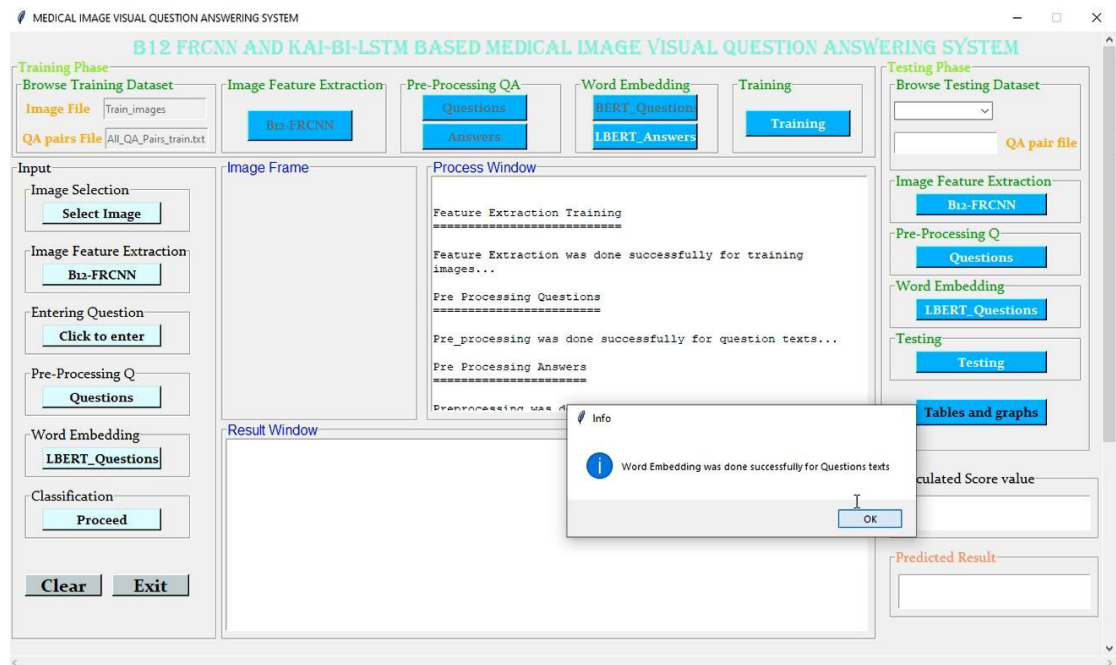


Fig 59 : Word Embedding for Question Dataset using BERT model

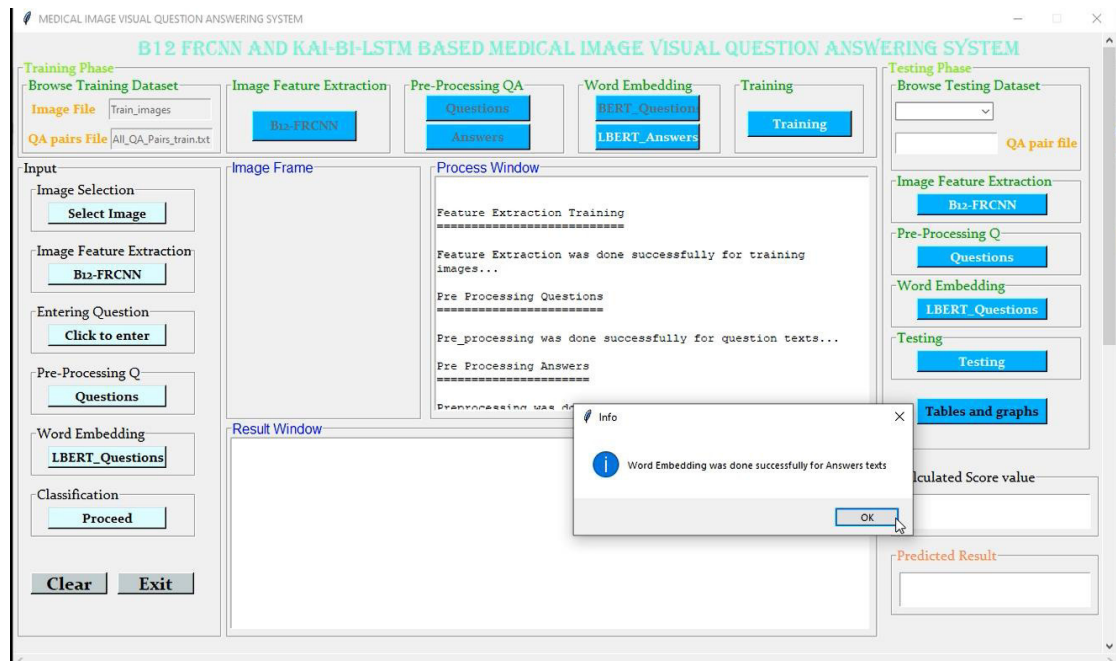


Fig 60 : Word Embedding for Answer Dataset using LBERT model

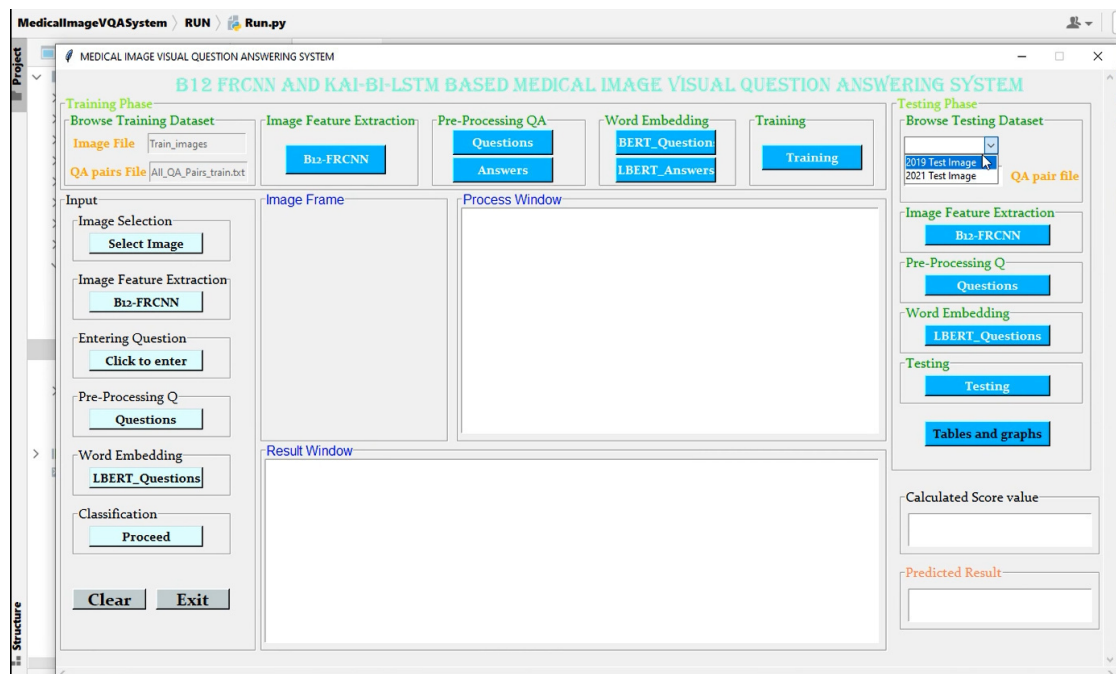


Fig 61 : Training the dataset for both visual and textual dataset

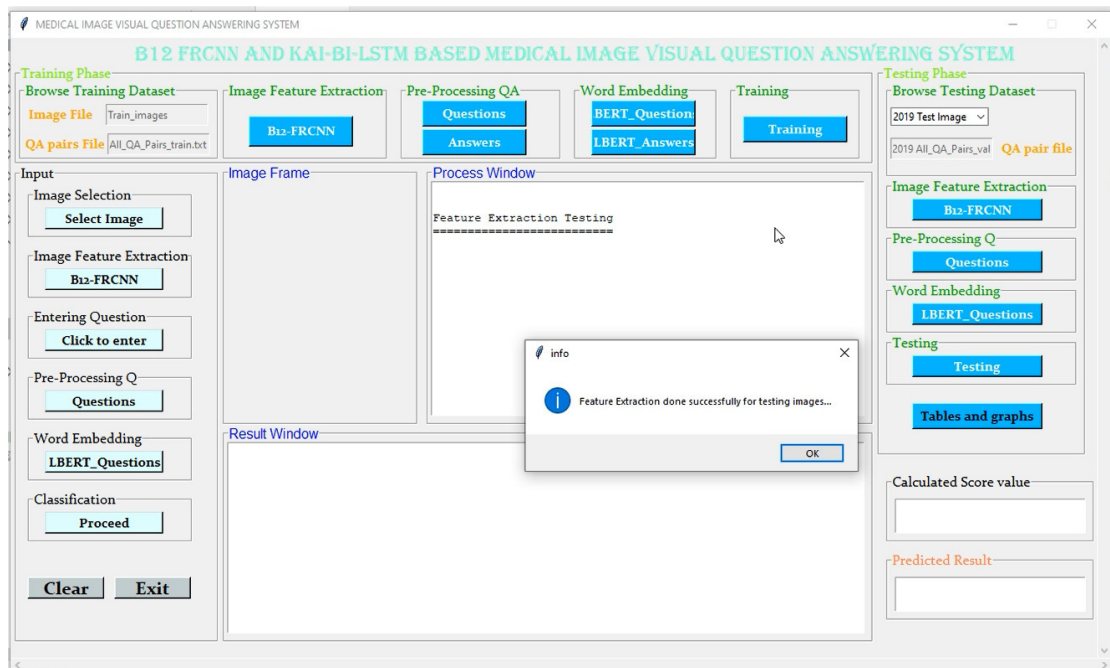


Fig 62 : Load the Testing dataset both Image and Question Answer pair

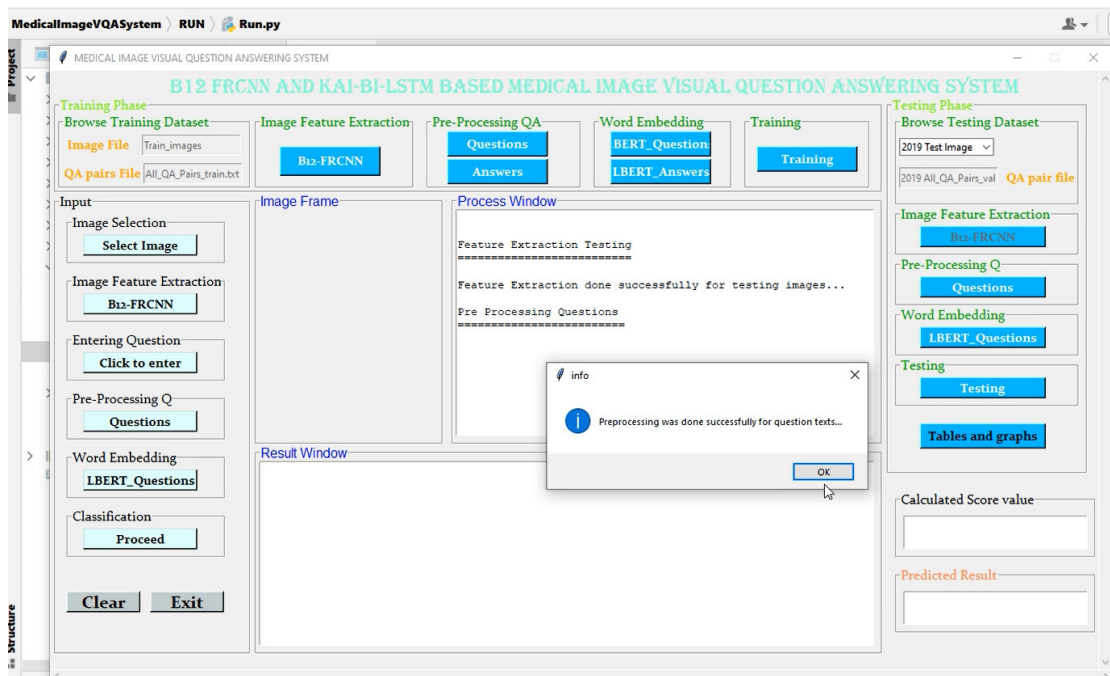


Fig 63 : Feature Extraction for Testing image dataset

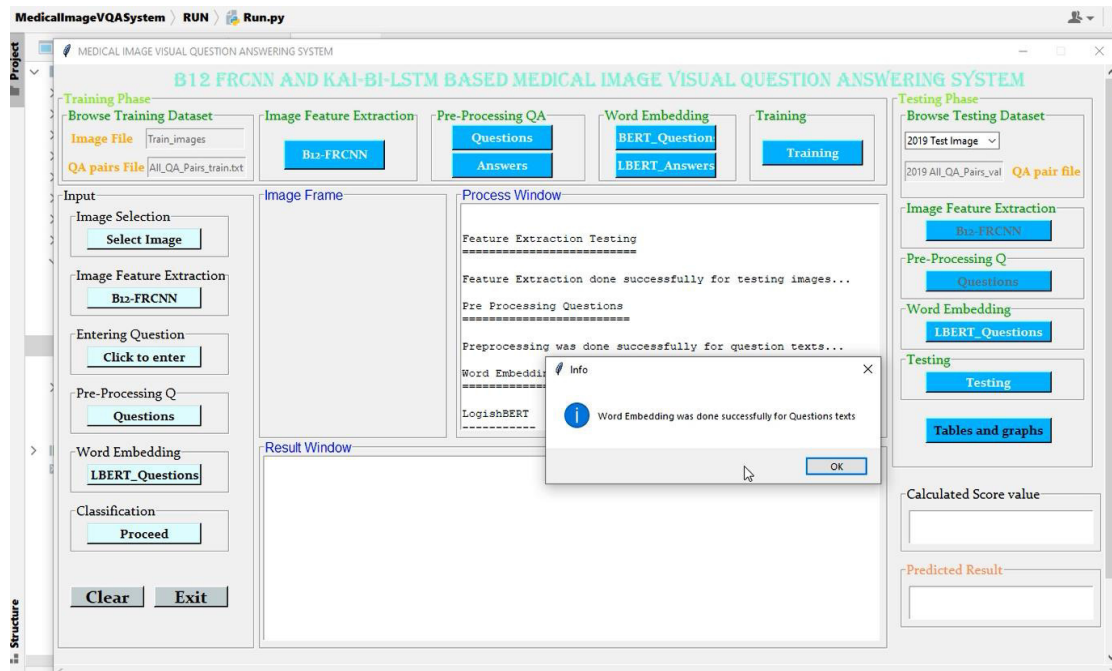


Fig 64 : Pre-processing for Question Datasets using LBERT model

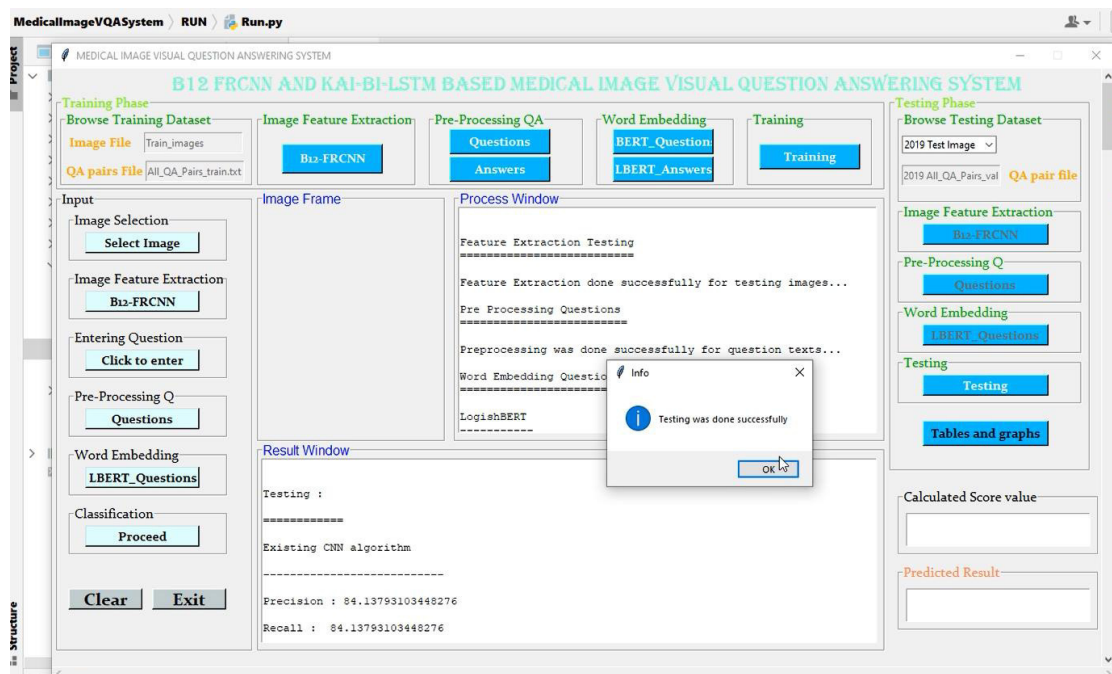


Fig 65 : Testing both visual and textual dataset

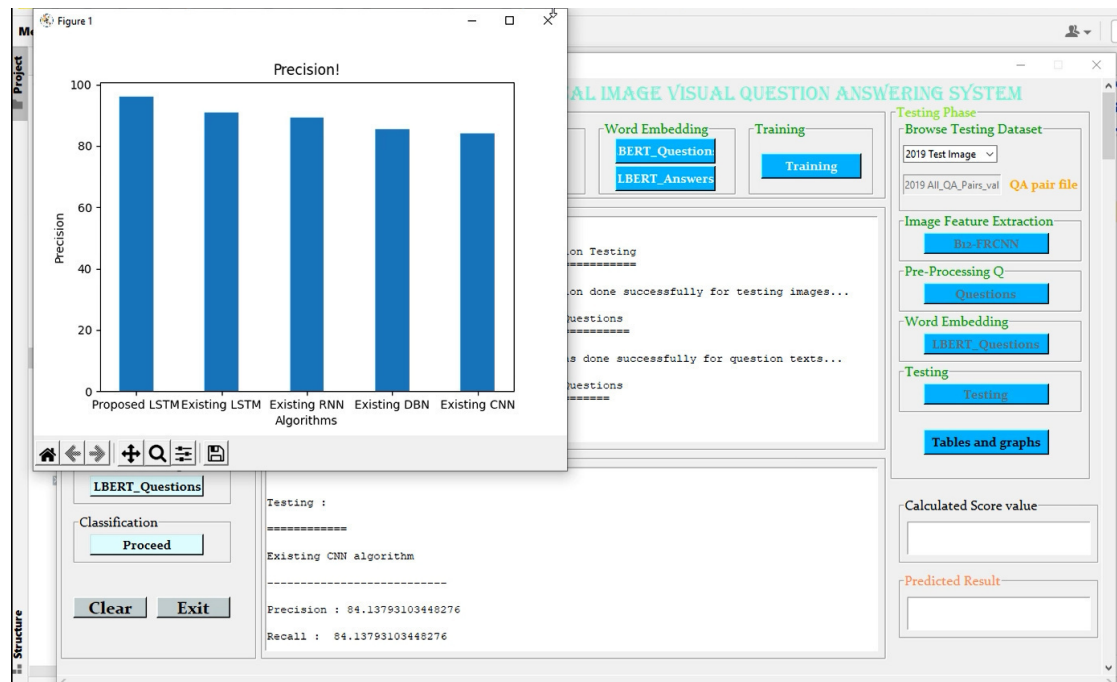


Fig 66- : Precision measure for proposed and existing models

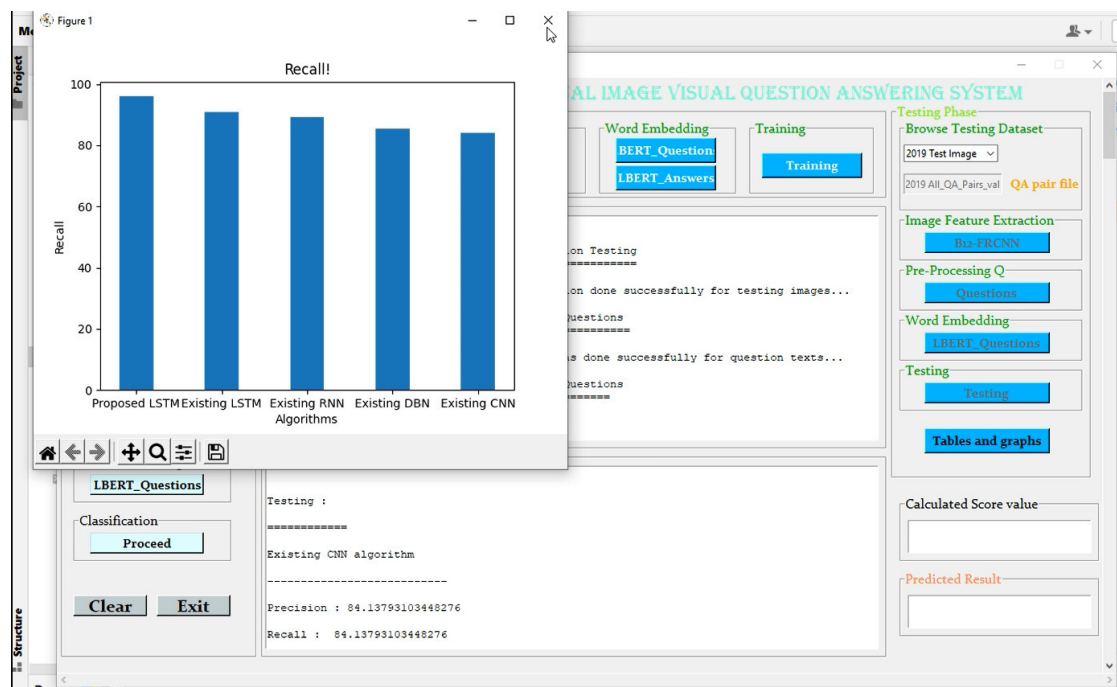


Fig 67 : Recall measure for proposed and existing models

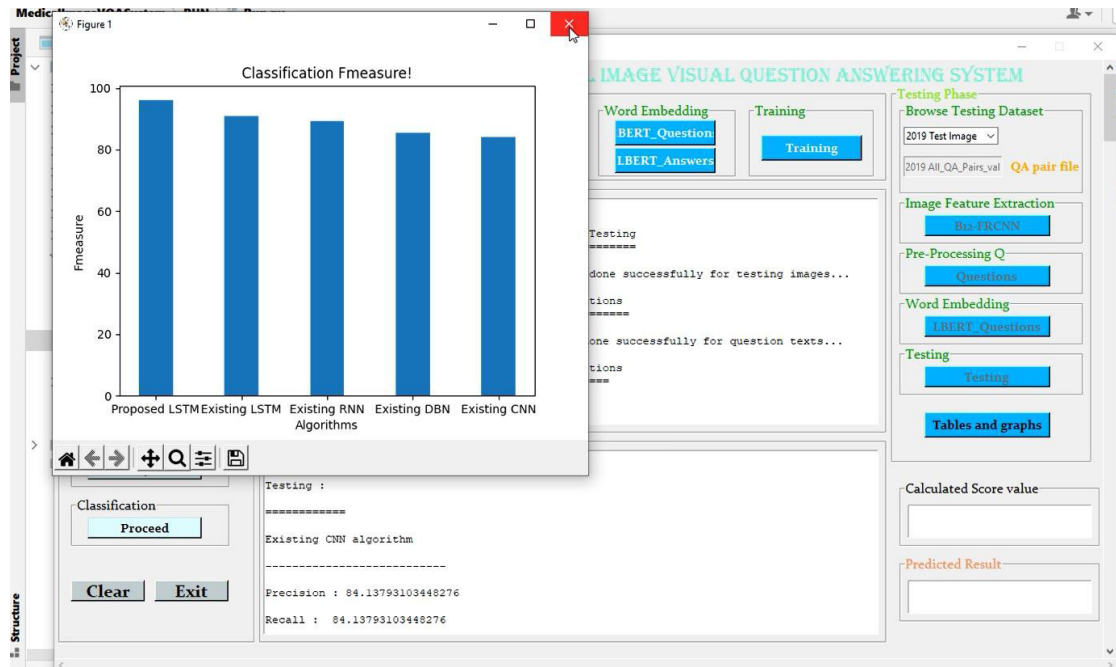


Fig 68 : Classification of FMeasure for proposed and existing models

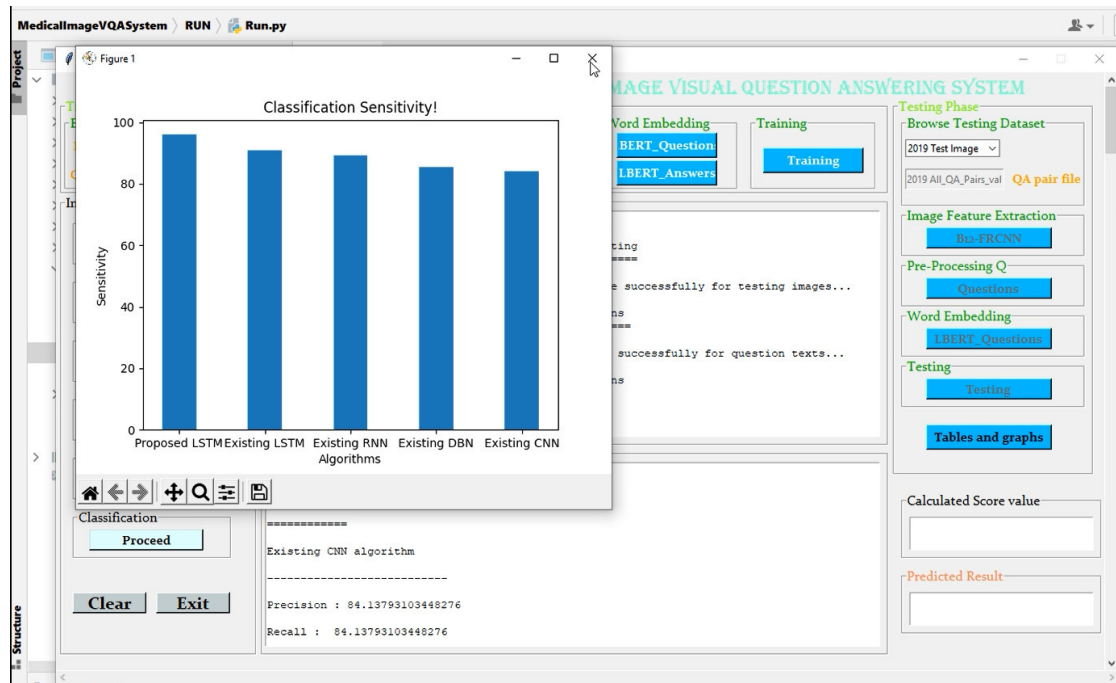


Fig 69 : Classification Sensitivity measure for proposed and existing models

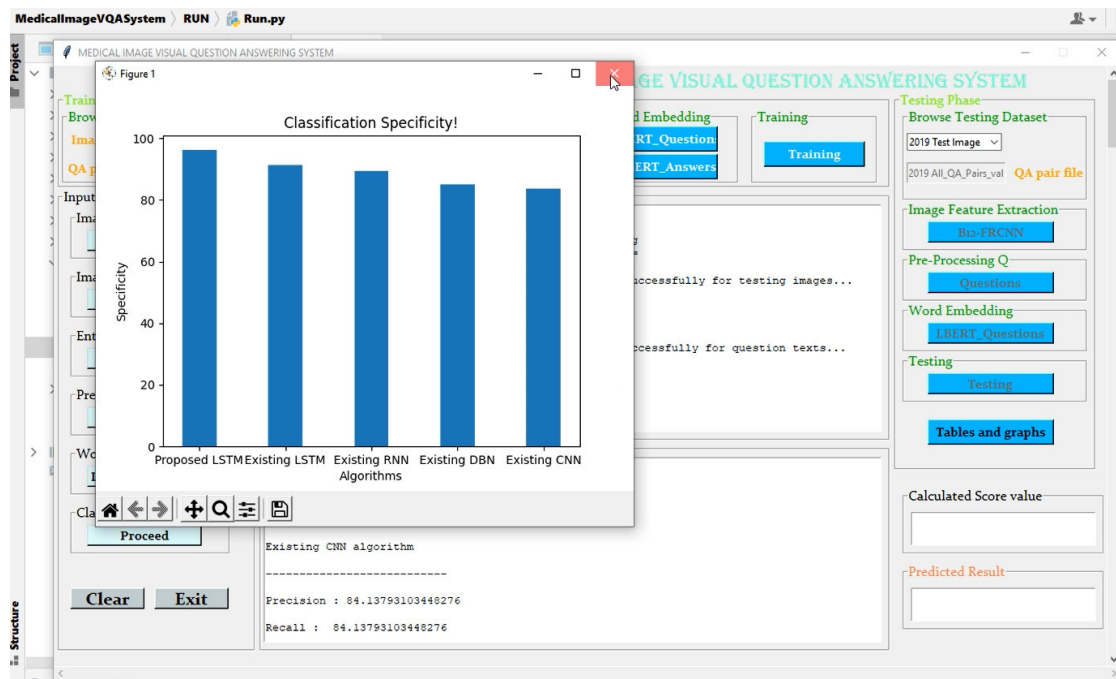


Fig 70 : Classification Specificity measure for proposed and existing models

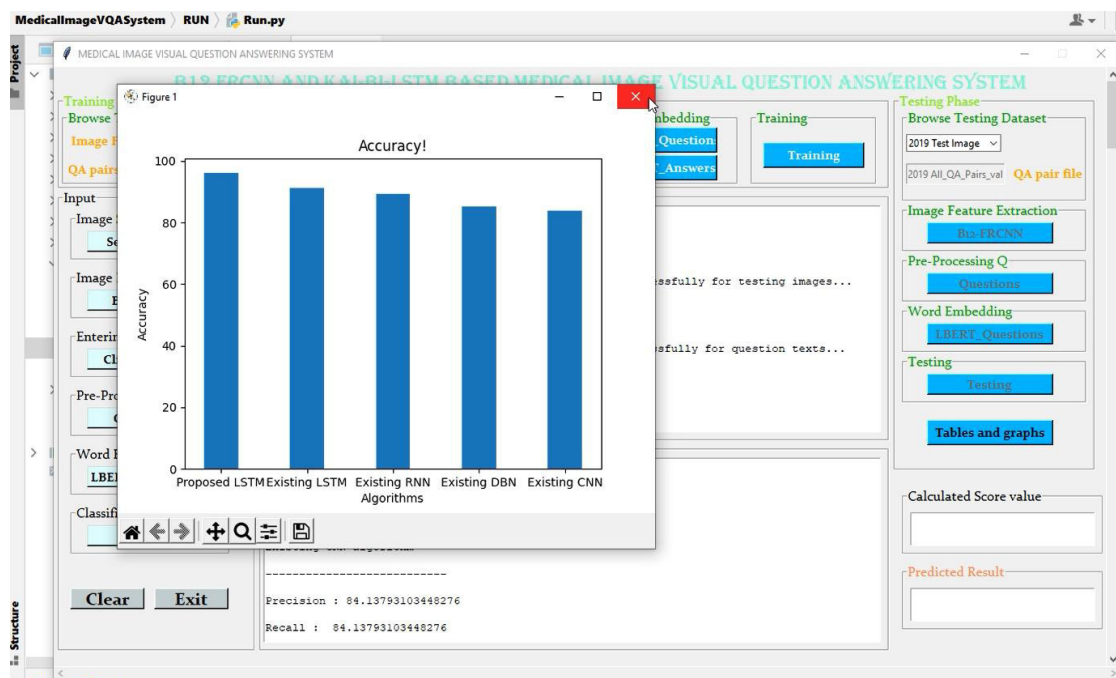


Fig 71 : True Positive Rate (TPR) measure for proposed and existing models

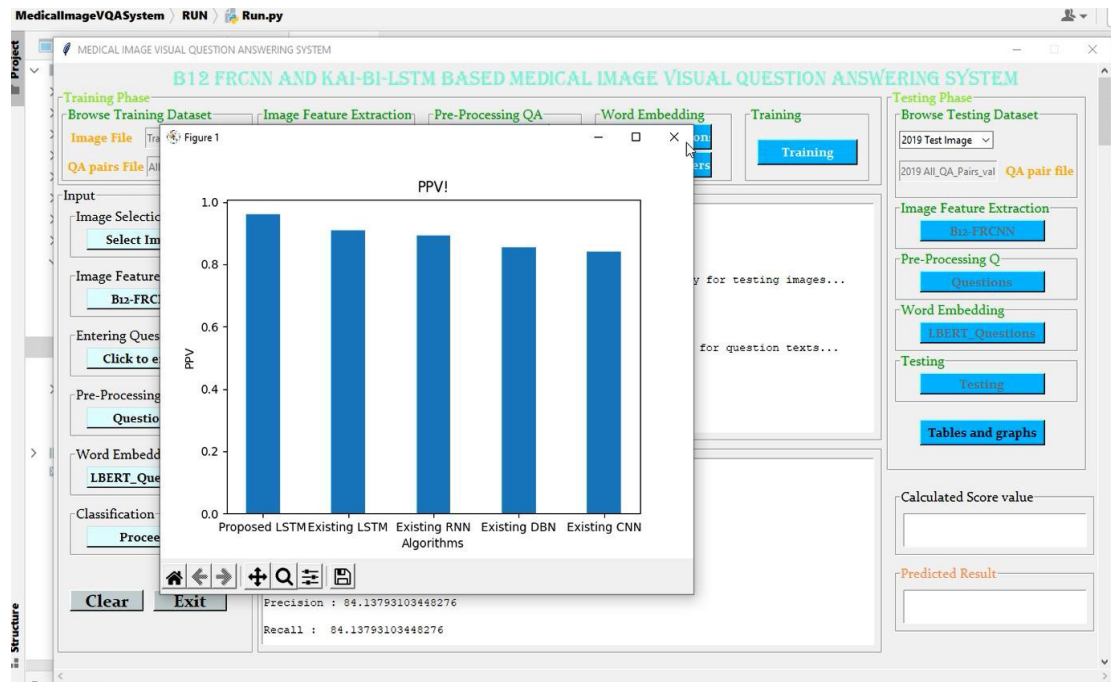


Fig 72 : Positive Predictive Value (PPV) measure for proposed and existing models



Fig 73 : Measure False Negative Rate (FNR) for proposed and existing models

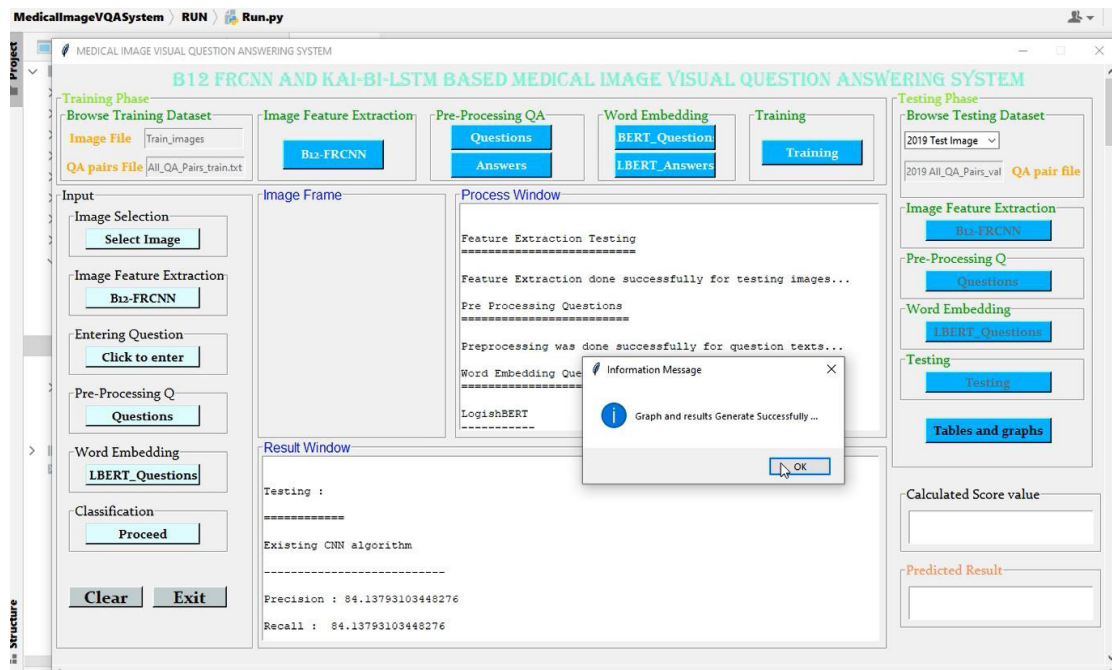


Fig 74 : To display Graphs and Tables

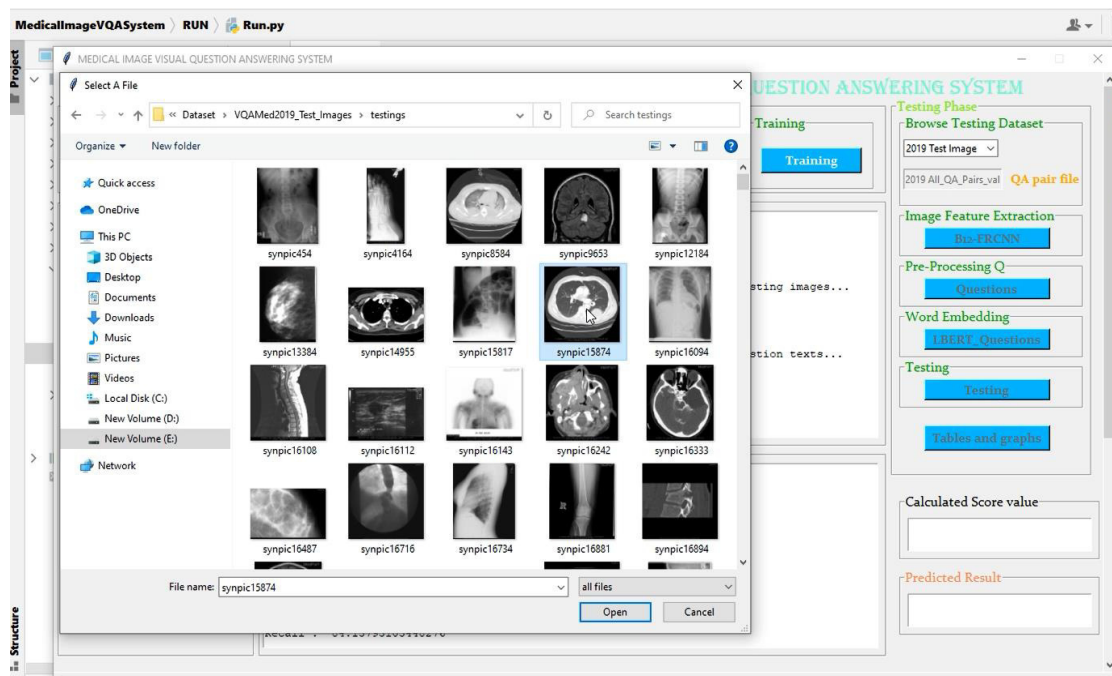


Fig 75 : Load the skeletal Image

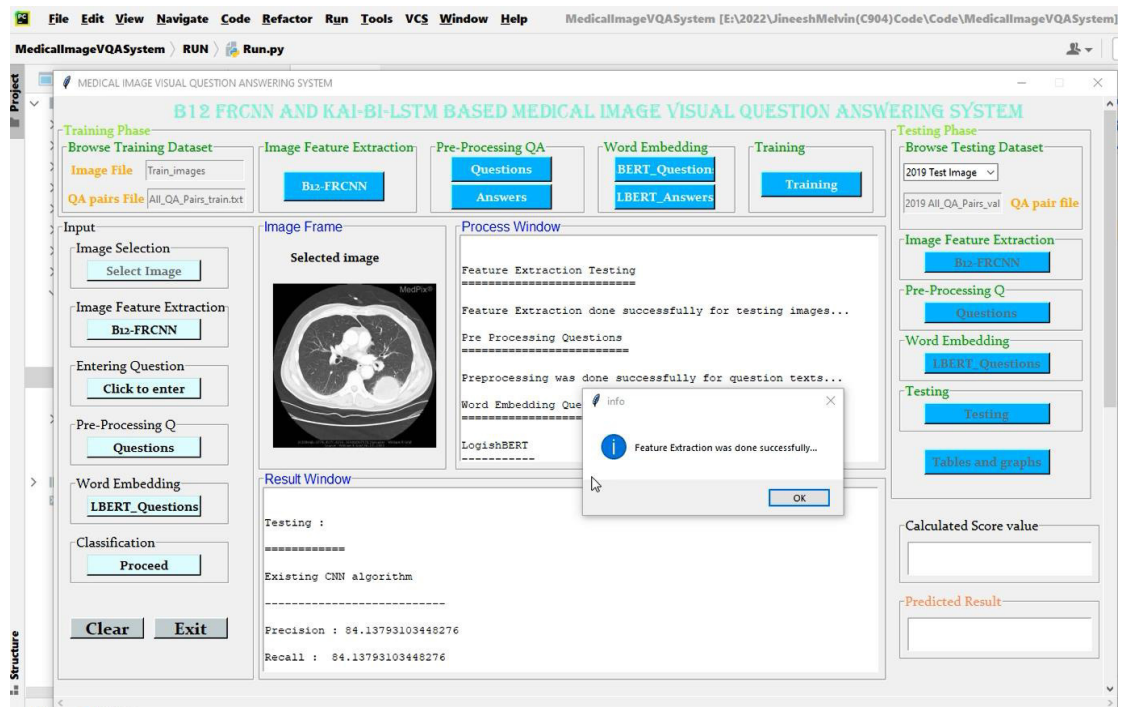


Fig 76 : Feature Extraction for Loaded Image using B12-FRCNN

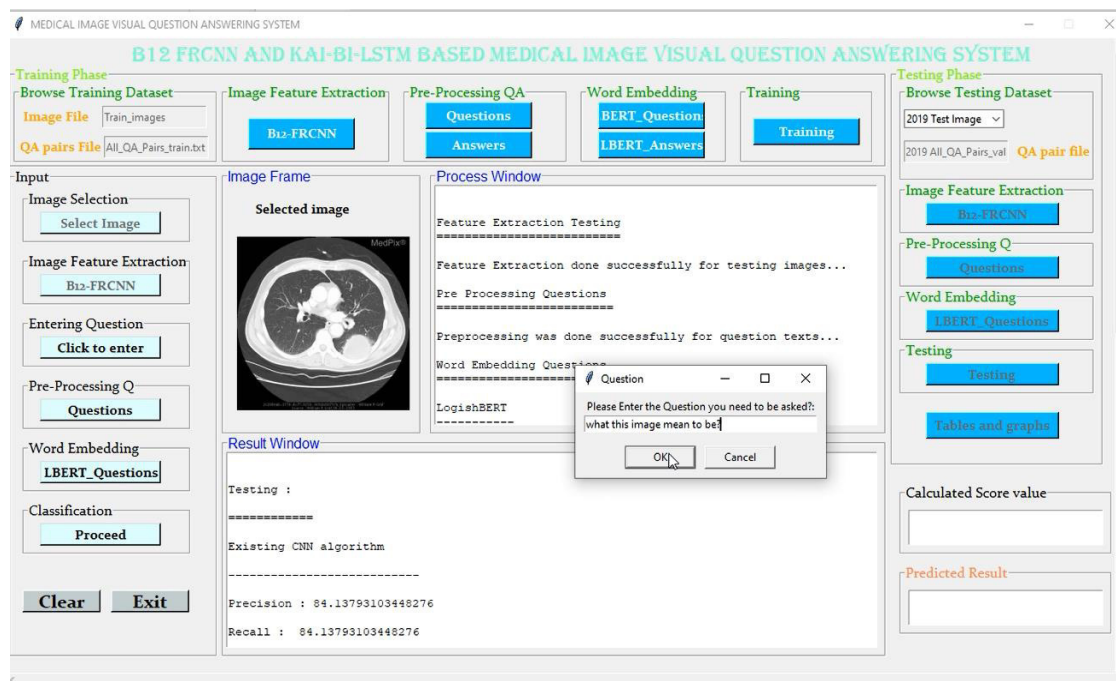


Fig 77 : Users Input Question

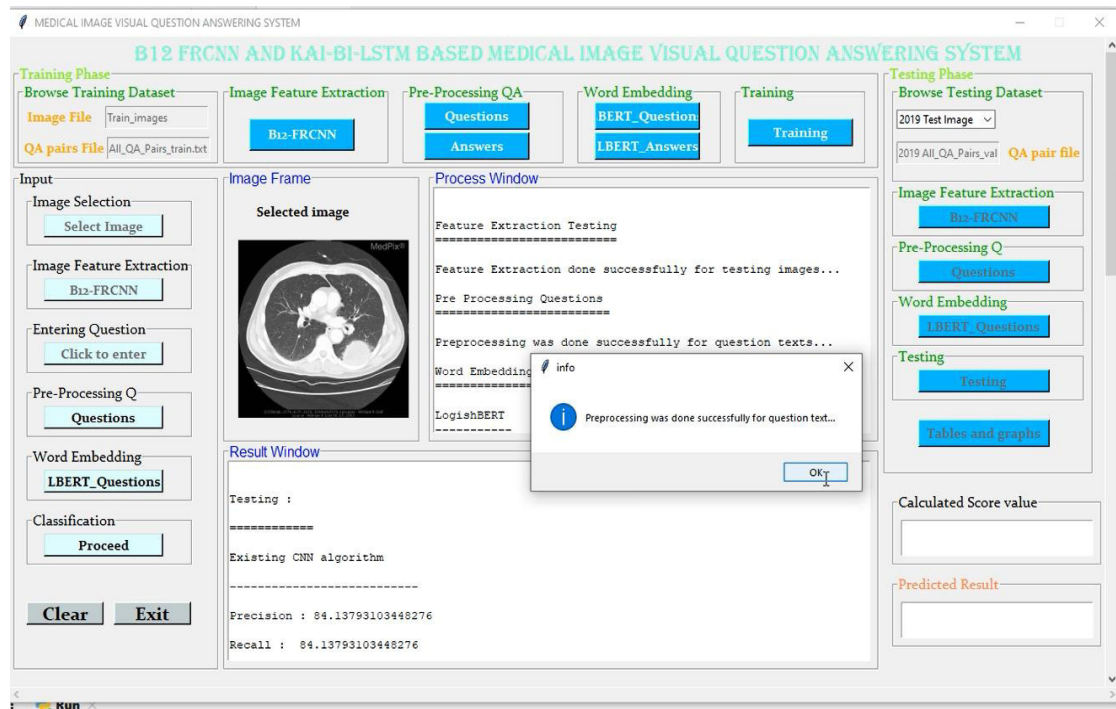


Fig 78 : Preprocess the Input Question

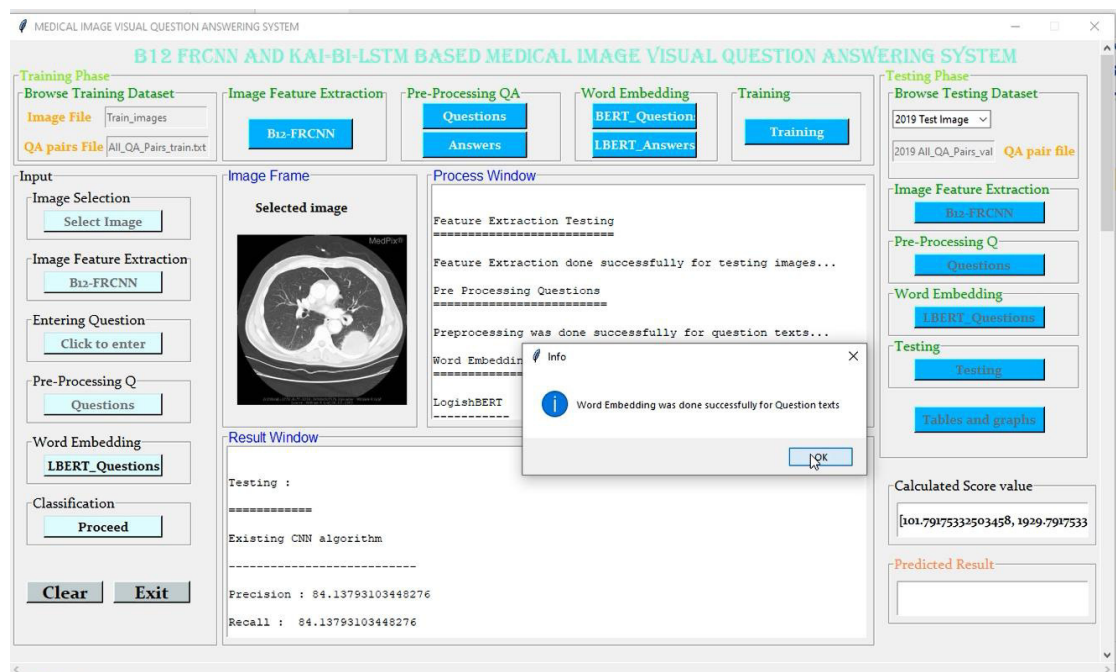


Fig 79 : Word Embedding for Input Question

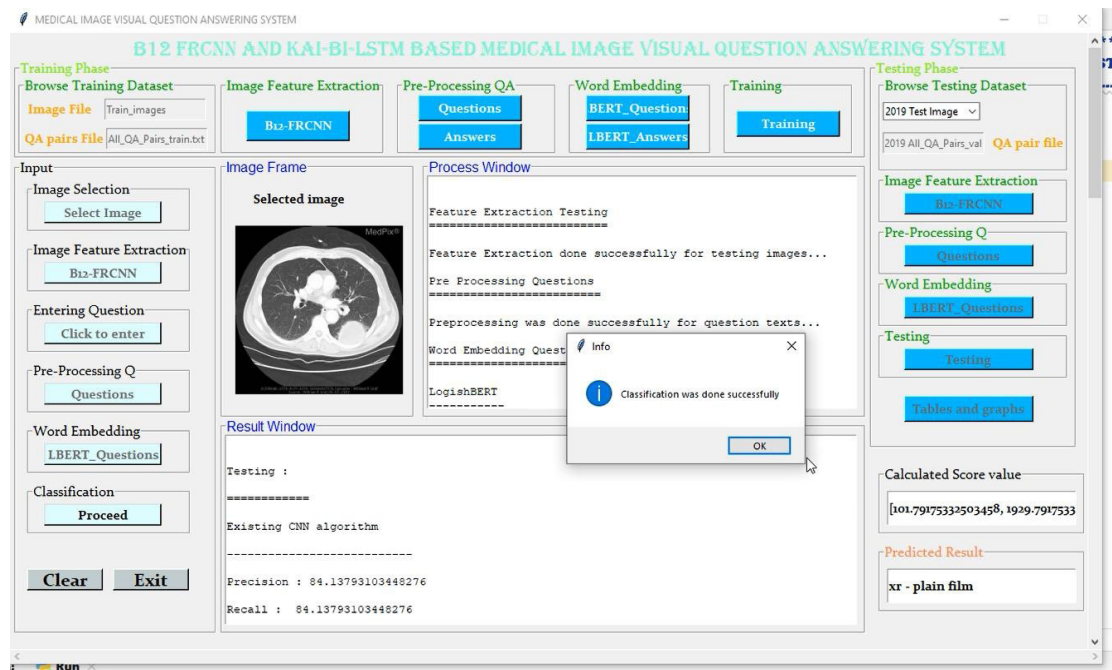


Fig 80 : Classification using Kai-BiLSTM model

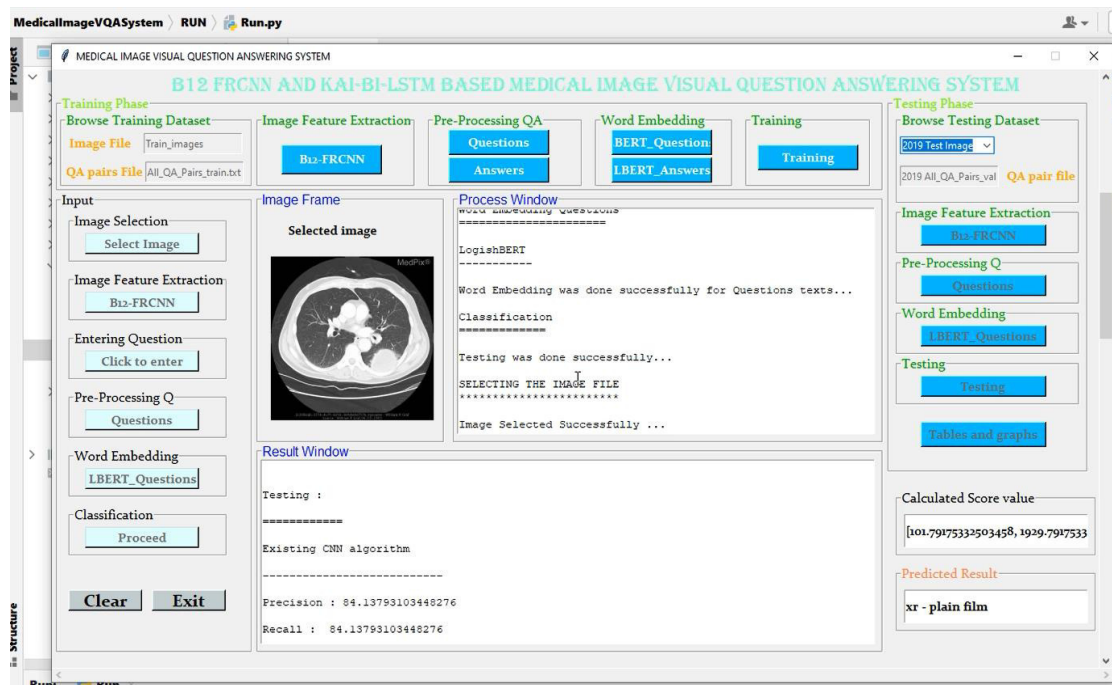


Fig 81 : Predicated answer with calculated score value for the answer

CONCLUSION AND FUTURE ENHANCEMENT



Visual Question Answering (VQA) systems hold significant promise for enhancing the capabilities of medical practitioners by offering automated assistance in the analysis of medical images. These technologies have the potential to improve medical diagnosis, practice efficiency, patient education, research and development, and remote telemedicine. With the rapid advancements in visual information processing and speech language processing, it is anticipated that VQA systems will become increasingly sophisticated and accurate. This improves patient consequences and the performance of medical treatments. As a result, VQA systems for medical imaging are expected to become indispensable tools for healthcare professionals in the near future.

The future adoption and efficacy of medical Visual Question Answering (VQA) systems will be determined by a number of factors, including the adequacy and caliber of medical Visual Question Answering (VQA) datasets, the creation and assessment of medical VQA models, and the assimilation and application of these systems within clinical environments. Addressing present dataset limits is critical, demanding the development of large and diverse medical VQA datasets that cover a wide range of modalities, symptoms, queries, and responses. Such initiatives are crucial for fostering advancements in medical VQA technology and its seamless integration into healthcare practices.

The B12FRCNN model is employed to extract visual feature knowledge, while BiLSTM is utilized for extracting textual feature information. The Kai-BiLSTM approach is then applied for data classification, achieving an impressive accuracy of 96.9% compared to other existing models. This advancement significantly enhances the quality of the visual question-answering system, enabling healthcare assistants to perform their tasks more efficiently. Notably, this model is adaptable to any dataset and yields reliable accuracy.

Our experimentation involved training and testing on CLEF Image Retrieval and Classification Task 2019, ImageCLEF 2020, and ImageCLEF 2021 datasets. Moving forward, leveraging additional datasets can further improve accuracy, potentially achieving 100%, and elucidating solutions within images to enhance users'

understanding of various scan reports. With access to large datasets, the model can be optimized for rapid accessibility, facilitating expedited analysis.

Future enhancements for visual question answering systems in the healthcare domain may include, Continuously gather and curate larger, more diverse, and comprehensive datasets specific to medical imaging. This includes datasets covering various modalities, conditions, anatomical structures, and abnormalities. Further refine existing models such as B12FRCNN and BiLSTM by fine-tuning their parameters and architectures to better suit the intricacies of medical imaging data. This could involve optimizing hyperparameters, exploring novel architectures, or incorporating domain-specific knowledge. Explore advanced feature extraction techniques tailored to medical images, such as attention mechanisms, graph-based methods, or self-supervised learning. These techniques can help capture more nuanced information from images and improve the performance of VQA systems. Investigate methods for effectively integrating data from diverse modalities, including imaging data, clinical text, and patient metadata. Fusion strategies such as late fusion, early fusion, or attention-based fusion can enhance the understanding of complex medical scenarios. Develop techniques for domain adaptation to mitigate the domain gap between training and deployment environments. This involves transferring knowledge from existing datasets to new domains, such as different hospitals or imaging protocols, to ensure the generalizability of VQA systems. Enhance the explainability and interpretability of VQA systems by incorporating mechanisms to provide reasoning or justification for the generated answers. This can improve trust and acceptance among healthcare professionals by making the decision-making process more transparent. Conduct rigorous clinical validation studies to assess the real-world performance and utility of VQA systems in clinical settings. Collaborate with healthcare professionals to ensure that the systems meet their needs and integrate seamlessly into existing workflows. Develop scalable and accessible VQA solutions that can be deployed across different healthcare settings, including hospitals, clinics, and remote areas with limited resources. Consider factors such as computational efficiency, ease of deployment, and user-friendly interfaces. Implement mechanisms for continual learning to adapt VQA systems over time as new data becomes available. This allows the models to stay up-to-date with evolving medical knowledge and adapt to changes

in clinical practices. Address ethical and regulatory considerations surrounding the deployment of VQA systems in healthcare, including patient privacy, data security, bias mitigation, and compliance with regulatory standards such as HIPAA. Ensure that the systems adhere to ethical guidelines and safeguard patient interests.

REFERENCE



REFERENCE

1. Bazi, Yakoub, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. 2023. "Vision–Language Model for Visual Question Answering in Medical Imagery" *Bioengineering* 10, no. 3: 380.
2. Claudio Filipi Goncalves dos Sants, Felype de Castro Bastos, Ana Claudia Akemi Matsuki de Faria et al. "Visual Question Answering: A Survey on Techniques and Common Trends in Recent Literature", 05 June 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-3015858/v1]
3. Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Yang Feng, Jin Hao, Junhui Lv, Zuozhu Liu. "Parameter-Efficient Transfer Learning for Medical Visual Question Answering", *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.
4. Zhihong Lin, Donghao Zhang, Qingyi Tac, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. "Medical visual question answering: A survey", 2022.
5. Lu S, Liu M, Yin L, Yin Z, Liu X, Zheng W. "The multi-modal fusion in visual question answering: a review of attention mechanisms". *PeerJ Comput Sci.* 2023 May 30;9:e1400. doi: 10.7717/peerj-cs.1400. PMID: 37346665; PMCID: PMC10280591.
6. V. Kodali and D. Berleant, "Recent, Rapid Advancement in Visual Question Answering: a Review," 2022 IEEE International Conference on Electro Information Technology (EIT), 2022, pp. 139-146.
7. Sruthy Manmadhan and Binsu C. Kavoov, "Visual question answering:a state-of-the-art review," *Artif. Int. Rev.*, vol. 53, pp. 5705-5745, 2020, https://doi.org/10.1007/s10462-020-09832-7.
8. Himanshu Sharma and Anand Singh Jalal. "A survey of methods, datasets and evaluation metrics for visual question answering. *Image and Vision Computing*", 116:104327, 2021.
9. Charulata Patil and Manasi Patwardhan. "Visual question generation: The state of the art". *ACM Comput. Surv.*, 53(3), may 2020.

10. Yeyun Zou and Qiyu Xie. "A survey on VQA: Datasets and approaches". In 2020 2nd International Conference on Information Technology and Computer Application (ITCA). IEEE, dec 2020.
11. Fuji Ren and Yangyang Zhou, "CGMVQA a new classification and generative model for medical visual question answering", IEEE Access, vol. 8, pp. 50626-50636, 2020.
12. Lubna A, Saidalavi Kalady and Lijiya A, "MoBVQA a modality based medical image visual question answering system", TENCON 2019 - 2019 IEEE Region 10 Conference IEEE, 17-20 October 2019, Kochi, India, 2019.
13. Fazal Muhammad, Ziaul Haq Abbas, Ghulam Abbas and Lei Jiao, "Decoupled downlink-uplink coverage analysis with interference management for enriched heterogeneous cellular networks", IEEE Access, vol. 4, pp. 6250-6260, 2016.
14. Dhruv Sharma, Sanjay Purushotham and Chandan K Reddy, "MedFuseNet an attention based multimodal deep learning model for visual question answering in the medical domain", Scientific Reports, vol. 11, no. 1, pp. 1-18, 2021.
15. Shengyan Liu, Xuejie Zhang, Xiaobing Zhou and Jian Yang, "BPI-MVQA a bi-branch model for medical visual question answering", BMC Medical Imaging, vol. 22, no. 1, pp. 1-19, 2022.
16. Yakoub Bazi 1, Mohamad Mahmoud Al Rahhal 2, Laila Bashmal 1 and Mansour Zuair 1 "Vision–Language Model for Visual Question Answering in Medical Imagery", Bioengineering 2023.
17. Li, L.; Lei, J.; Gan, Z.; Liu, J. Adversarial "VQA: A New Benchmark for Evaluating the Robustness of VQA Models". In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2022–2031.
18. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding", arXiv 2019, arXiv:1810.04805.
19. He K, Zhang X, Ren S, Sun J. "Deep residual learning for image recognition". In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016;770–778.

20. Srinivasan K, Garg L, Datta D, Alaboudi AA, Jhanjhi NZ, Agarwal R, Thomas AG. "Performance comparison of deep cnn models for detecting driver's distraction". *CMC-Comput Mater Continua*. 2021;68(3):4109–24.
21. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. "Learning phrase representations using rnn encoder-decoder for statistical machine translation", *arXiv preprint arXiv: 1406. 1078*, 2014.
22. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. "Biobert: a pre-trained biomedical language representation model for biomedical text mining". *Bioinformatics*. 2020;36(4):1234–40.
23. Devlin J, Chang M-W, Lee K, Toutanova K. "Bert: pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv: 1810. 04805*, 2018.
24. Peng Y, Liu F, Rosen MP. "Umass at imageclef medical visual question answering (med-vqa) 2018 task". In *CLEF (Working Notes)*, 2018.
25. Zhejiang University at CLEF Image Retrieval and Classification Task 2019 'Visual Question Answering in the Medical Domain. 2019'.
26. Kornuta T, Rajan D, Shivade C, Asseman A, Ozcan AS. "Leveraging medical visual question answers with supporting facts", *arXiv preprint arXiv: 1905. 12008*, 2019.
27. Liao Z, Wu Q, Shen C, Van Den Hengel A, Verjans J. "Aiml at Healthcare Image QA 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering", 2020.
28. Al-Sadi A, Hana'Al-Theiabat, Al-Ayyoub M. "The inception team at Healthcare Image QA 2020: Pretrained vgg with data augmentation for medical vqa and vqg". In *CLEF (Working Notes)*, 2020.
29. Zhan L-M, Liu B, Fan L, Chen J, Wu X-M. "Medical visual question answering via conditional reasoning". In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020;2345–2354.
30. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: "Healthcare Image QA: Overview of the medical visual question answering task at CLEF Image Retrieval and Classification Task 2019". In: *CLEF (Working Notes)* (2019)

31. Y. I. Jinesh Melvin, S. Gawade and H. Palivela, "Visual Question Answering using Data Mining Techniques for Skeletal Scintigraphy in medical domain - VQADMSS," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, 2021, pp. 859-863, doi: 10.1109/ICAIS50930.2021.9395936.
32. Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. "Towards vqa models that can read". In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8309–8318, 2019.
33. Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís Gomez, Marçal Rusinol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. "Scene text visual question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)", October 2019.
34. Chao Yang, Su Feng, Dongsheng Li, Huawei Shen, Guoqing Wang, and Bin Jiang. "Learning content and context with language bias for visual question answering". In 2021 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6, July 2021.
35. Dirk V'ath, Pascal Tilli, and Ngoc Thang Vu. "Beyond accuracy: A consolidated tool for visual question answering benchmarking". In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 114–123, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
36. A. Farinhas, A. T. Martins, and P. Q. Aguiar. "Multimodal continuous visual attention mechanisms". In 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pages 1047–1056, Los Alamitos, CA, USA, Oct 2021. IEEE Computer Society.
37. Corentin Kervadec, Grigory Antipov, "Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to"? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2776–2785, 2021.

38. D. Teney, E. Abbasnejad, and A. van den Hengel. “Unshuffling data for improved generalization in visual question answering”. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 1397–1407, Los Alamitos, CA, USA, Oct 2021. IEEE Computer Society.
39. Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton Van den Hengel, and Qi Wu. “Structured multimodal attentions for textvqa. IEEE Transactions on Pattern Analysis and Machine Intelligence”, 2021.
40. Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z Pan, Zonggang Yuan, and Huajun Chen. “Zero-shot visual question answering using knowledge graph”. In International Semantic Web Conference, pages 146–162. Springer, 2021.
41. Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. “Medical visual question answering via conditional reasoning”. In Proceedings of the 28th ACM International Conference on Multimedia, MM ’20, page 2345–2354, New York, NY, USA, 2020. Association for Computing Machinery.
42. Vatsal Goel, Mohit Chandak, Ashish Anand, and Prithwijit Guha. “Iq-vqa: Intelligent visual question answering”. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, Pattern Recognition. ICPR International Workshops and Challenges, pages 357–370, Cham, 2021. Springer International Publishing.
43. S. Whitehead, H. Wu, H. Ji, R. Feris, and K. Saenko. “Separating skills and concepts for novel visual question answering”. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5628–5637, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society.
44. Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. “Debiased visual question answering from feature and sample perspectives”. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, 2021.
45. Rajat Koner, Hang Li, Marcel Hildebrandt, Deepan Das, Volker Tresp, and Stephan Günnemann. Graphhopper: “Multi-hop scene graph reasoning for visual question answering”. In Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam Barnaghi, Armin Haller, Mauro Dragoni,

-
-
- and Harith Alani, editors, *The Semantic Web – ISWC 2021*, pages 111–127, Cham, 2021. Springer International Publishing.
46. Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. “Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering”. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3601–3605, Florence, Italy, July 2019. Association for Computational Linguistics.
 47. Junjie Wang, Yatai Ji, Jiaqi Sun, Yujiu Yang, and Tetsuya Sakai. “Mirtt: Learning multimodal interaction representations from trilinear transformers for visual question answering”. pages 2280–2292, 01 2021.
 48. Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. “Contrastive Pre-training and Representation Distillation for Medical Visual Question Answering Based on Radiology Images”, pages 210–220. 09 2021.
 49. Kushal Kafle, Robik Shrestha, Scott Cohen, Brian Price, and Christopher Kanan. “Answering questions about data visualizations using efficient bimodal fusion”. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 1498–1507, 2020.
 50. Binh X Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. “Coarse-to-fine reasoning for visual question answering”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4558–4566, 2022.
 51. Haifan Gong, Ricong Huang, Guanqi Chen, and Guanbin Li. “Sysu-hcp at Healthcare Image QA 2021: A data-centric model with efficient training methodology for medical visual question answering”. *Proceedings* <http://ceur-ws.org> ISSN, 1613:0073, 2021.
 52. Anwen Hu, Shizhe Chen, and Qin Jin. “Question-controlled text-aware image captioning”. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3097–3105, 2021.
 53. Yibing Liu, Yangyang Guo, Jianhua Yin, Xuemeng Song, Weifeng Liu, Liqiang Nie, and Min Zhang. “Answer questions with right image regions: A visual attention regularization approach”. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(4):1–18, November 2022.
-
-

-
-
54. Leonard Salewski, A. Sophia Koepke, Hendrik P. A. Lensch, and Zeynep Akata. CLEVR-x: “A visual reasoning dataset for natural language explanations”. In *xxAI- Beyond Explainable AI*, pages 69–88. Springer International Publishing, 2022.
 55. Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. “VLMo: Unified vision-language pre-training with mixture-of-modality-experts”. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
 56. Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. “Greedy gradient ensemble for robust visual question answering”. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021.
 57. Zujie Liang, Haifeng Hu, and Jiaying Zhu. “LPF: A language-prior feedback objective function for de-biased visual question answering”. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, July 2021.
 58. Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, and Erik Learned-Miller. “Passage retrieval for outside-knowledge visual question answering”. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, July 2021.
 59. Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R. Manmatha. “Latr: Layout-aware transformer for scene-text vqa”. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16527–16537, June 2022.
 60. Emanuele Vivoli, Ali Furkan Biten, Andres Mafla, Dimosthenis Karatzas, and Lluís Gomez. “Must-vqa: Multilingual scene-text vqa”. In *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, page 345–358, Berlin, Heidelberg, 2023. Springer-Verlag.
 61. C. Kervadec, T. Jaunet, G. Antipov, M. Baccouche, R. Vuillemot, and C. Wolf. “How transferable are reasoning patterns in vqa?”. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4205–4214, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society.

-
-
62. Aisha Urooj Khan, Hilde Kuehne, Chuang Gan, Niels da Vitoria Lobo, and Mubarak Shah. “Weakly supervised grounding for vqa in vision-language transformers”. 2022.
 63. Herring W, Learning radiology: “Recognizing the basics. Elsevier Health Sciences”, 2015.
 64. Y. I. Jinesh Melvin, Sushopti Gawade, Hemant Palivela, "Feature Extraction from Radiology Images for Visual Question Answering System Using CNN and BiLSTM Model", Recent Innovations in Computing, vol.832, pp.317, 2022.
 65. Novelline RA and Squire LF, “Squire’s fundamentals of radiology”. La Editorial, UPR, 2004.
 66. Jones J, Normal abdominal x-ray. Case study, Radiopaedia.org (Accessed on 01 Feb 2024) <https://doi.org/10.53347/rID-34067>.
 67. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Müller, H., Lungren, M.P., 2018. “Overview of ImageCLEF 2018 medical domain visual question answering task”, in: CLEF (Working Notes).
 68. Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D., 2018. “A dataset of clinically generated visual questions and answers about radiology images”. Scientific Data 5, 1–10.
 69. Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H., 2019. Healthcare Image QA: Overview of the medical visual question answering task at CLEF Image Retrieval and Classification Task 2019, in: CLEF2019 Working Notes, CEUR-WS.org, Lugano, Switzerland.
 70. Kovaleva, O., Shivade, C., Kashyap, S., Kanjaria, K., Wu, J., Ballah, D., Coy, A., Karargyris, A., Guo, Y., Beymer, D.B., et al., 2020. “Towards visual dialog for radiology”, in: Proceedings of the 19th SIG-BioMed Workshop on Biomedical Language Processing, pp. 60–69.
 71. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P., 2020. “PathVQA: 30000+ questions for medical visual question answering”. arXiv preprint arXiv:2003.10286 .
 72. Ben Abacha, A., Datla, V.V., Hasan, S.A., Demner-Fushman, D., Müller, H., 2020. “Overview of the Healthcare Image QA task at ImageCLEF 2020:

-
- Visual question answering and generation in the medical domain”, in: CLEF 2020 Working Notes, CEUR-WS.org, Thessaloniki, Greece.
73. Ben Abacha, A., Sarrouiti, M., Demner-Fushman, D., Hasan, S.A., Müller, H., 2021. “Overview of the Healthcare Image QA task at ImageCLEF 2021: Visual question answering and generation in the medical domain”, in: CLEF 2021 Working Notes, CEUR-WS.org, Bucharest, Romania.
 74. Liu, S., Ou, X., Che, J., Zhou, X., Ding, H., 2019. “An Xception GRU model for visual question answering in the medical domain”, in: CLEF (Working Notes).
 75. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D., 2017. “Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering”, in: Conference on Computer Vision and Pattern Recognition (CVPR).
 76. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. “Microsoft COCO: Common objects in context”, in: European conference on computer vision, Springer. pp. 740–755.
 77. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S., 2019. “MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs”. arXiv preprint arXiv:1901.07042 .
 78. Liu, B., Zhan, L.M., Wu, X.M., 2021a. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), “Medical Image Computing and Computer Assisted Intervention – MICCAI 2021”, Springer International Publishing, Cham. pp. 210–220.
 79. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M., Golia-Pernicka, J., Heckers, S.H., Jarnagin, W.R., McHugo, M.K., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J., “A large annotated medical
-

- image dataset for the development and evaluation of segmentation algorithms”, 2019.
80. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017b. “ChestX-Ray8: Hospital-scale chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases”, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 3462–3471.
 81. Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S., 2019. “CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data”.
 82. Girshick, “R. Fast r-cnn”. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
 83. Ren, S.; He, K.; Girshick, R.; Sun, J. “Faster r-cnn: Towards real-time object detection with region proposal networks”. IEEE Trans. Pattern Anal. Mach. Intell. 2016, 39, 1137–1149.
 84. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
 85. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, “A. You only look once: Unified, real-time object detection”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
 86. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. “Ssd: Single shot multibox detector”. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
 87. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. “Rich feature hierarchies for accurate object detection and semantic segmentation”. arXiv, 2014; arXiv:1311.2524.

88. He, K.; Zhang, X.; Ren, S.; Sun, J. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition". arXiv, 2014; arXiv:1406.4729.
89. Wang, X.; Shrivastava, A.; Gupta, A. "A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection". arXiv, 2017; arXiv:1704.03414.
90. Ren, S.; He, K.; Girshick, R.; Sun, J. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". IEEE Trans. Pattern Anal. Mach. Intell. **2017**, 39, 1137–1149.
91. Joseph, R.; Santosh, D.; Ross, G.; Ali, "F. You Only Look Once: Unified, Real-Time Object Detection". arXiv, 2015; arXiv:1506.02640.
92. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. "SSD: Single shot multibox detector". arXiv, 2016; arXiv:1512.02325.
93. Simonyan, K.; Zisserman, "A. Very Deep Convolutional Networks for Large-Scale Image Recognition". arXiv, 2014; arXiv:1409.1556.
94. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, "Inception-ResNet and the Impact of Residual Connections on Learning". arXiv, 2016; arXiv:1602.07261.
95. Abacha, Asma Ben and Gayen, Soumya and Lau, Jason J and Rajaraman, Sivaramakrishnan and Demner-Fushman, Dina. "NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain" (2018).
96. Hochreiter, S., Schmidhuber, J., 1997. "Long short-term memory". Neural Computation 9, 1735–1780.
97. Jiang, M., Chen, S., Yang, J., Zhao, Q., 2020. "Fantastic answers and where to find them: Immersive question-directed visual attention", in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 2977–2986.
98. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S., 2019. "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs". arXiv preprint arXiv:1901.07042 .

-
-
99. Jung, B., Gu, L., HaradaAl-Sadi, T., 2020. bumjun_jung at Healthcare Image QA 2020: “VQA model based on feature extraction and multi-modal feature fusion”, in: CLEF 2020 Working Notes.
 100. K. Verma, H., Ramachandran S., S., 2020. “HARENDRAKV at Healthcare Image QA 2020: Sequential VQA with attention for medical visual question answering”, in: CLEF 2020 Working Notes.
 101. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D. & Parikh, D. “Making the v in vqa matter: Elevating the role of image understanding in visual question answering”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6904–6913 (2017).
 102. Allaouzi, I. & Ahmed, M. B. “Deep neural networks and decision tree classifier for visual question answering in the medical domain”. In CLEF (Working Notes) (2018)
 103. C. Wang, H. Yang, and C. Meinel, “Image captioning with deep bidirectional LSTMs and multi-task learning,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 14, no. 2, 2018.
 104. T. Liu, S. Yu, B. Xu, and H. Yin, “Recurrent networks with attention and convolutional networks for sentence representation and classification,” *Applied Intelligence*, vol. 48, no. 10, pp. 3797–3806, 2018.
 105. Y. Yang, W.-T. Yih, and M. C. Wikiqa, “A challenge dataset for open-domain question answering,” in Proceedings of the Conference Empirical Methods Natural Language Processing, Lisbon, Portugal, September 2015.
 106. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
 107. B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, “Simple baseline for visual question answering,” 2015, arXiv:1512.02167
 108. T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear CNN models for finegrained visual recognition,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1449–1457.
 109. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2001, pp. 311–318.

110. Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. 2015. “ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering”. arXiv e-prints (Nov. 2015), arXiv:1511.05960

PUBLICATIONS



Visual Question Answering using Data Mining Techniques for Skeletal Scintigraphy in medical domain - VQADMSS

Mr. Jinesh Melvin YI
 Department of Computer Engineering
 Pillai College of Engineering,
 New Panvel, Navi Mumbai, India
yijmelvin@mes.ac.in

Dr. Sushopti Gawade
 Department of Computer Engineering
 Pillai College of Engineering,
 New Panvel, Navi Mumbai, India
sgawade@mes.ac.in

Dr. Hemant Palivela
 Head of AI and ML
 eClerx Services Ltd.
 Mumbai, India
hemant.datascience@gmail.com

Abstract—Understanding about the medical images of patients is a very tedious task. Doctors should convey their patient through the image of the questions asked by the patient. Large amounts of labeled data are required for training in traditional approaches for VQA (Visual Question Answering). Also, the description of clinic trial text in English and in multilingual contexts is one of the challenges in the medical field. To present the clarification about the images, doctors are required to provide the related images. It is better for comparison with the patient's previous report and current report. This paper contributes to solve the problems related to VQA for better description of the image and accuracy related images through the answer of the questions, also to make it easy to convey the users with any kind of images. Question answer process should be more descriptive for easy to understand and traceable. This system helps to identify the types of images which are captured by any scanner. The better accuracy is a visualization method which projects the answers as a baseline that shows the corresponding region with various colors, which is easier to note the answers in visual method for the appropriate questions. This proposed framework focuses on Radiology image for Skeletal Scintigraphy to transform and generate a model using Data Mining Techniques. This system suggests that the effective medical Visual Question answering techniques is better to assist doctors in clinical analysis and diagnosis. This also will help the hospital services to grow the medical domain.

Keywords—Radiology images, Classification models, Generative models, Transformer, Visual Question Answering

I. INTRODUCTION

Medical domain field is rapidly growing with different tools and techniques to improve the benefits of patients, researchers and Clinicians. Past few years, research in computer science and medical science have been developing intelligent tools for supporting medical decision making. Different software's was designed by various vendors to help the doctors, patients and clinicians. Researchers are also keen to provide new techniques with the help of technology to help society. The difficulties faced by patients are to understand about their health and body conditions. To know exact information about body and health communication between Doctor and patient communication is

very important. While discussing health conditions, patients may or may not understand the terms which are used by doctors or clinicians. Various soft computing systems have been successfully developed in health care professionals to support patients about this. Many radiology centers are available around for patient's benefits with the latest techniques and tools. In case of radiology images, patients need to consult some doctors regarding their health. Consulting people will raise many questions about the radiology image to clarify their health conditions. Manually it is difficult to answer all the queries of patients, so this research proposes automatic answers about the queries asked by patients.

A. Overview of Radiology image

Radiology is an imaging technology that is used to diagnose and treat disease and it is a branch of medicine. Radiology has two different areas, diagnostic radiology and interventional radiology. Diagnostic radiology is used to visualize the structure inside the body. The specialist will use integration of these images. It helps to show all the symptoms and also to monitor how well the body responds to the treatment patient is receiving for specific disease. It helps to visualize different illnesses, such as colon cancer, heart disease and breast cancer etc. The most common types of diagnostic radiology exams include CT (Computed Tomography) also known as CAT (Computerized Axial Tomography). It includes CT angiography, Fluoroscopy with upper GI, magnetic resonance imaging (MRI) scan and magnetic resonance angiography (MRA) scan, mammography, bone scan, thyroid scan, plain x-ray, PET images, PET scan, PET-CT, ultrasound. As the user enters the input image, this system will compare with the database and convey it to the user with proper information for further process.

Interventional Radiology imaging is helpful to doctors when inserting wires, catheters and other small instruments and tools into our body; it is a smaller incision cut. It often involved treating blockages, problems in veins, problems in uterus, back pain, liver problems and kidney problems. Proposed system helps to identify Interventional Radiology images. This

research deals with both types of radiology images related to different diseases.

B. Question Answer Processing

The proposed system VQADMSS supports question answering mechanisms to help patients as shown in Fig 1. The huge amount of data should be trained for predicting an accurate output, once the raw data gets trained with large amounts of questions with subjective and descriptive answers. The large datasets have a large number of variables to compute the resource to process them. The effective amount of data reduction without losing any data is one of the biggest challenges in this system. The reduction of data helps to reduce the machine effort and generate the new model to increase the speed up the performance and generalize the process of learning to the machine. After the feature extraction process the quality datasets get stored in a new location and data mining techniques will help to classify with a random forest algorithm to predict the output answer for the question from the image will be accurate. Test the machine with an image and ask different questions, so that the data mining classification algorithm gets processed and predicts the proper output and it asks whether to predict the related image for further clarification of input images and questions.

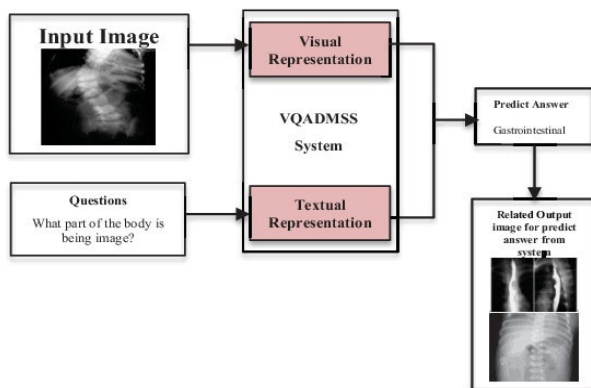


Fig. 1. Block diagram of QA System.

The various medical assists launched in the medical domain to improve and ease access, but Visual Question Answering (VQA) is somewhat different techniques also it gives benefits to any kind of patient. There are lots of ways to study about our health conditions without any expert guidance. Computer techniques played major roles in the medical domain and it grew in medical health and hospital services. The system which helps the patients will give more clarity in any type of radiology image using VQA. Our paper focused on Skeletal Scintigraphy, about bone marrow, bone cancer, density, infection, osteonecrosis, osteoporosis etc. It assists for the same to patients, also it helps with a multilingual system for illiterate people. It gives more predicted values during the Question Answering session.

Many Different questions considered by any patients have not satisfying the methods to solve their comprehensive problem, in QA System diagnosis CGMVQA model generates

capabilities to turn the complex problem to simple problems, which includes supervised and question generation, with data augmentation on images and tokenization on texts. It minimizes the parameter of the multi-head self-attention transformer to cut the computational cost down, and also add different kinds of embedding together to deal with text [2]. PICO is a framework for Data Mining in clinical Trial Text which transforms for classification and question answering tasks. It is mainly focused on detecting and annotation of information about PICO elements. The characteristics of PICO is Population, Intervention, Comparator and Outcome. It allows creation with the aim to support systematic reviews of semi automation using NLP method. It contributes to the sentence prediction task, training the data from different tasks and integrating. Prepared the dataset using training and testing subsets that fit the SQuAD format and merged both the dataset in order to check the correct answer for PICO questions and flexibility of general purpose of answering for questions on the basis of SQuAD [1].

Closed-end and Open-end are the two important components in medical visual question answering via conditional reasoning. There are multiple questions arise from users by two types of questions mainly underlined. In closed-end tasks only the answers should be limited like yes or no but in open-end tasks the answers should be free texts. Image is given as the input and output will be the answers in two forms. Question condition reasoning module to guide the modulation of multimodal fusion features [3].

II. RELATED WORK

(Deepak Gupta, Swati Suman and Asif Ekbal, 2010) designed a new system with three different techniques such as Question Segregation techniques, then the second stage of this system is to integrate the question segregation model with hierarchical deep multi-model neural network and with the impact of question segregation which compare the performance of hierarchical deep multimodal network with question segregation and without question segregation. SVM algorithm is used as a base classifier to extract feature vectors for question segregation. This system produces two different types of input model which are yes/no and others. Examined this system with RAD and CLEF18 datasets in [4].

In the Question Answer with Deep Learning a survey has an Automatic Question Answering system takes place in three different groups, such as deep neural network, dynamic memory network and rational networks. In this the datasets to be two different categories one is textual and other one is in visual. Finally it evaluates the matrix for information retrieval system and automatically text generation. Deep Learning in neural networks which predict a solution for a task and it concludes with previous experience. The system gets processed between two sets of data that concern input and output to solve a task. The two main common architectures used for survey, they are Convolution Neural Network (CNN) and Recurrent Neural Network (RNN). To solve the sequence tag is to be used by dynamic memory networks also for classification problems, sequence to sequence tasks and question answering tasks. To compute the relation without the need to be learned in which recurrent neural networks learn to capture sequential dependency and Convolution Neural Network (CNN) learn spatial dependencies.

As on Image CLEF 2020: Multimedia Retrieval in Lifelogging Medical, Nature and Internet application follows four different tasks in which the first task is all about life log, it cares videos, images and other sources about daily activities understanding, retrieval and summarization. Then the next to analyze the caption, predict the tuberculosis and answer the questions of medical images which mean by Medical tasks. The segmenting and labeling collection of coral image for 3D modeling is a coral task and the final task is about the web user interfaces to detect and recognize the problem which were addressed from hand drawn websites for generating automatic code [6].

Domain Specific Question Answering system is based on a knowledge graph which is also meant by causality knowledge graph and it indicates the process of question answer. This relationship redefined and extracted the knowledge graph of doctor's. Once the question answer pair generated from concept knowledge graph to train the machine learning models for retrieval. Concept Knowledge Graph is used to display the schema of data, which improves the completeness. Doctors are invited to set some rules to convert classes into triples. So that the experts have enough knowledge about the relationship of the concepts in the medical domain. Instance knowledge graphs are specific entities like disease or drugs etc. Hybrid method is used to retrieve the answers. The system tries two different models for every Natural Language Question. The traditional method is used by support vector machines and sequence to sequence deep learning models. In which SVM classifies for discrete and continuous variables, but sequence to sequence deals with text. Structure query based models intend to translate the Natural Language Questions submitted by users into SPARQL Query statements. The system will first execute the word segmentation using the jieba word segmentation tool. Medical dictionaries tools help to improve its reliability. The system converts questions into vectors and CNN model classifies the questions to find best-fit then it will execute the SPARQL query and return the results to the users. Question Answer presents subgraphs with corresponding Natural Language sentences. It makes the answer reasonable. It helps users to get more knowledge or information about his or her questions [7].

To extract the visual feature from CNN (Convolutional Neural Network) & Global Average Pooling strategy. It captured the medical images using training datasets. The BERT model plays a major role to encode the question which is raised with semantic features. It scores the accuracy of 0.624 also the image CLEF 2019 has selected as a trained image. Bidirectional Encoder Representation from Transformers has been released two size BERT_{BASE} and BERT_{LARGE}. BERT_{BASE} has 12 layers in the Encoder stack but BERT_{LARGE} has 24 layers in the Encoder stack. BERT is an Encoder stack of transformer architecture through the network, which uses self-attention. The different stages of these systems extract the feature from the imported image, parallelly it encoded the semantic question using the BERT model, then it forward to feature fuse with co-attention mechanism. It helps to avoid irrelevant information for different regions. Image CLEF 2019 visual question answering in the medical domain has used multi-modal factorized bilinear pooling. The performance of this model is very efficient and effective for VQA, because of the combination with multi model features. But Bilinear is outperform, also its traditional linear modes for VQA. It gives limited capacity and low performance [9]. The final model is to predict the answers with

the accuracy score of 0.624, also the perfect mismatch score for this system is 0.644[8].

Two different classification models used in the medical VQA framework named MAML and CDAE. This model is used to initialize the model weights for image feature extraction. MAML represents Model-Agnostic Meta-Learning, training the model with a dataset for MAML by manually reviewing more number of question and answer pairs. The images from the dataset get separated into three categories such as head, chest and abdomen, also it gets divided into subgroups based on question answer pairs corresponding to the images. From all of these it categorized into 9 classes. Unlabeled images are used to train CDAE and encoder by feeding them before input images which use Gaussian noise to corrupt [11]. After training both MAML and CDAE this system used trained weight MEVF image feature extraction components in VQA framework, then finetune the whole VQA model using a train set of VQA-RAD dataset, which makes genuine comparison to [12]. The VQA accuracy is computed as the percentage for open-ended and close-ended for both MAML and CDAE models from scratch and finetuning. From this analysis finetuning has more accuracy than scratch.

III. PROPOSED SYSTEM

In Visual Question Answering for Skeletal Scintigraphy system collect any type of Radiology image there is no limits of image selection. Diagnostic and Interventional are the two different types of Radiology image. It's an imaging technology to diagnose and treat diseases. This system focuses on skeleton scintigraphy which is considered as the study of bone scans. It helps to detect the fractures in bones, cancers in bones, insufficiency fractures, affection of bone joints and many others. This helps to answer the Question for all types of bones in the human body from skull to foot, such as Long bone, Short bone and irregular bones. The most common types of diagnostic radiology are computed tomography which we called CT scan also it named as Computerized axial tomography (CAT) scan, the other type of radiology images are magnetic resonance imaging (MRI) and magnetic resonance angiography (MRA), plain X-ray, positron emission tomography also known as PET imaging, PET scan or PET-CT and ultrasound. The other type of Radiology image is Interventional radiology, it helps the doctors to insert catheters, wires, tools and some small instruments into the human body. Doctors can detect and treat the diseases directly through a scope which is mentioned as a camera with open surgery, some of the Interventions are Nuclear Medicine Imaging, PET, X-ray, and Ultrasound.

Question answering process should be more descriptionable for easy to understand and traceable. This helps any kind of patient and doctors to get a clear view about the images, also it clear the maximum doubts by giving proper description. This system helps to identify the types of images captured by the medical imaginary tools like whether the image is Diagnostic Radiology or Interventional Radiology. Visualization the answers as baseline and it shows the corresponding regions based on the question. Suppose the doctors or patients want to know the explanation of the particular area from the image, so the system identifies the proper region which enquiry by the users by mapping the baseline as in visualization mode as mentioned in Fig. 3. As the description of answers predicted from the images

which the questions asked by the users as mentioned in Table 1, also this system helps to retrieve the related images from the answer as reference to the user, so that it is more understandable and convenient to the users for further treatments. It seems like the answers from images and the reference images predicted from the answers.

IV. SYSTEM DESIGN AND METHODOLOGY

The classification technique is the better suggestion in Machine Learning to develop this system. The huge amount of radiology medical images to be collected from various sources as a dataset for implementing the system, in which we use a set of images for training and test the ratio for supervised learning is 70:30. The 70% of dataset will be used to train the machine with different answers from the image like type of images, Question types, Name of the image, image position and which part of the organ. Before training the dataset, the image extraction module will extract the feature image from the normal image. The Inception Resnet V2 model is used for image feature extraction as it is a type of Convolutional neural network, which trains more than millions of images from the Image Net database. Also it classifies the images into 'n' number of object categories. It gives good performance and decreases the training cost. This model enables connection shortcuts to speed up network training. Bidirectional layer input models run in two ways, one from the past to future and other vice-versa. It helps to extract the question features, with LSTM. Bi-LSTM generates the representation of questions [4]. Once the image gets trained with the above ratio, then it is classified using machine learning algorithms. Random Forest is the best choice in supervised learning for classification the trained dataset with more accurate results. It has the direct relationship between the number of trees in the forest and the results will be in an efficient manner. Compared with other algorithms like SVM, Regression, K mean, the accuracy of RF is higher. RF constructs a decision tree from the given dataset for every sample and it will predict the results from every decision tree, then the voting will perform for every predicted result. The final prediction result is the most voted tree from the dataset. The predicted output will be the answer for the question asked by the user. Finally it fetches the related part of the image from the backend for more effectiveness or it refers the image from the answer for further treatments.

TABLE I. Sample Questions and answer pairs formulated from a single image, more than one medical related question asked from a given image.

Questions	Objective Answers	Subjective Answer	Organ	Image Type
What does the CT scan show?	left atrium	A large filling defect in the left atrium.	Fetal Skull	Diagnostic
Where is the anterior fontanel?	Top		Fetal Skull	Diagnostic
Is it normal?	Yes		Fetal Skull	Diagnostic

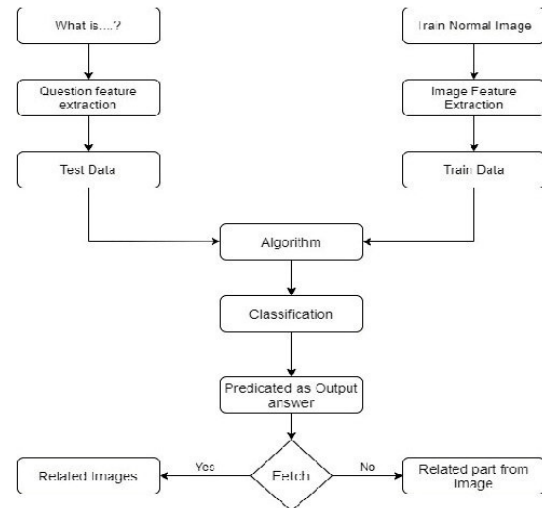


Fig. 2. Design of Proposed System.

V. USE CASE RELATED ON PROPOSED SYSTEM

The Organ shown in Fig 3 is Fetal Skull. It's an input image by the user, which has two types of view (i) Superior view and (ii) Lateral View. The result has Questions, Type of Questions, and Answers in Objective type and Subjective type, Name of the Organ and Image Type as shown in Table 1. The displayed image parts name will list out to the users in the same page as mentioned in Fig. 4, when the user clicks on any listed name it will show the part in the displayed image with base red line. The related image option will also be available on the same screen to get more clarity about the image.

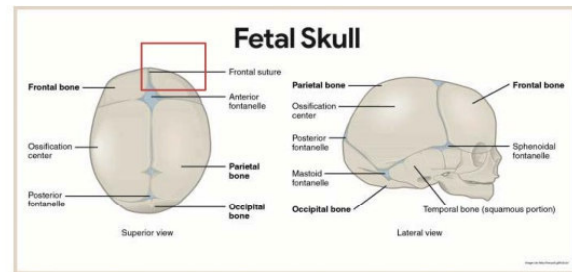


Fig. 3. Fetal Skull in Superior view and Lateral view [13]

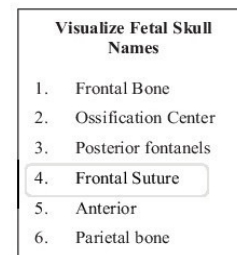


Fig. 4. List the names of Fetal Skull in Superior view

CONCLUSION

This system will explore the Answering for users question in visual mode for medical images, it will help users to take further procedure after getting proper results. The contribution of this system is to train the normal images and extract the feature from the image and it trains the data for classification. It predicts the proper answers in a descriptive manner with better accuracy. Also it will retrieve the related images for user's reference. Different algorithms used in survey papers which have less accuracy also only identifies the answers for questions. The cost of training set is very low when we compare with other models as we prefer Resnet v2 model and Bi-LSTM for image and question feature extraction. Use cloud technology for more accessible to the users where they can utilize this system in any remote area with more predictable as future enhancement.

REFERENCES

- [1] Lena Schmidt, Julie Weeds and Julian P. T. Higgins, "Data Mining in Clinical Trial Text: Transformers for Classification and Question Answering Tasks," University of Bristol, Bristol Medical School, 39 Whatley Road, BS82PS Bristol, UK, Jan 2020.
- [2] Fuji Ren, (Senior Member, IEEE), and Yangyang Zhou, "CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering," Faculty of Engineer, University of Tokushima, Tokushima 770-8506, Japan, March 6, 2020.
- [3] Li-Ming Zhan, Bo Liu, Lu Fan, "Medical Visual Question Answering via Conditional Reasoning", The Hong Kong Polytechnic University, October 2020.
- [4] Deepak Gupta, Swati Suman, Asif Ekbal, "Hierarchical deep multi-modal network for medical visual question answering", Department of Computer Science and Engineering, Indian Institute of Technology Patna, India, Sep 2020.
- [5] Martina Toshevska, Georgina Mirceva, Mile Jovanov, "Question Answering with Deep Learning: A Survey," Faculty of Computer Science and Engineering Ss. Cyril and Methodius University Skopje, Macedonia, 11 March 2020.
- [6] Bogdan Ionescu, Henning M'uller, Renaud P'eteri, "ImageCLEF 2020: Multimedia Retrieval in Lifelogging, Medical, Nature, and Internet Applications," University Politehnica of Bucharest, Bucharest, Romania, Aug 2020.
- [7] Ming Sheng, Anqi Li, Yuelin Bu, BNRist, "DSQA: A Domain Specific QA System for Smart Health Based on Knowledge Graph," DCST, RIIT, Tsinghua University, Beijing 100084, China, Aug 2020.
- [8] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu, "ImageCLEF 2019 Visual Question Answering in the Medical Domain," Zhejiang University, Hangzhou, China, Sep 2019.
- [9] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning M'uller, "VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019," Lister Hill Center, National Library of Medicine, USA, 2019 Sep.
- [10] Scibert: Pre-trained contextualized embeddings for scientific text. ArXiv, abs/1903.10676, Beltagy, L., Cohan, A., and Lo, K. (2019).
- [11] Jjj Binh D. Nguyen, Thanh-Toan Do, Binh X. Nguyen, Tuong Do, Erman Tjiputra, and Quang D. Tran, "Overcoming Data Limitation in Medical Visual Question Answering", AIOZ Pte Ltd, Singapore, 26 Sep 2019.
- [12] Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Nature (2018)
- [13] <https://i.pinimg.com/originals/66/01/ee/6601ee3b13a9e24a6aa53942fe1f8ce1.jpg> is accessed on 20 Jan 2021.
- [14] Asma Ben Abacha, Soumya Gayen, Jason J. Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman. 2018. NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain. In Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, Vol. 2125). CEUR- WS.org, Avignon, France.
- [15] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR. IEEE Computer Society, Salt Lake City, UT, USA, 6077–6086.
- [16] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," IEEE Access, vol. 6, pp. 9375–9389, 2018.
- [17] R. Wang, X. Liang, X. Zhu, and Y. Xie, "A feasibility of respiration prediction based on deep bi-LSTM for real-time tumor tracking," IEEE Access, vol. 6, pp. 51262–51268, 2018.

Novel approach to integrate various feature extraction techniques for the Visual Question Answering System with skeletal images in the healthcare sector

Jinesh Melvin Y I

Computer Engineering
Pacific Academy of Higher Education and Research University
Udaipur, India
jm3998@gmail.com

Sushopti Gawade

Computer Engineering
Mumbai University
Mumbai, India
sushoptikrishimitra@gmail.com

Mukesh Shrimali

Computer Engineering
Pacific Academy of Higher Education and Research University, Pacific Hills
Udaipur, India
Mukesh_shrimali@yahoo.com

Abstract— In the realm of medical science, one of the most challenging concepts to grasp is the Medical Imaging Query Response System. The comprehension and classification of the diverse representations of the human body require a significant degree of effort and expertise. Furthermore, it is imperative for users within the healthcare sector to rigorously validate the system. In the domain of human health, a plethora of imaging techniques, including MRI, CT, ultrasound, X-ray, PET-CT, and others, play a pivotal role in the identification of medical issues. These technologies are instrumental in supporting both patient engagement and clinical decision-making. However, the utilization of models, techniques, and datasets for processing textual and visual information introduces complexities that can at times impede the provision of pertinent clinical solutions. The overarching objective of the proposed approach is to conduct a comprehensive comparative analysis of various feature extraction methodologies for both visual and textual information within the Visual Question Answering (VQA) system, focusing on human skeletal images. This endeavor is aimed at enhancing the VQA system's performance with newer datasets and addressing any limitations inherent in existing models. In addition, this research initiative seeks to enable researchers to identify and optimize novel methods that enhance the accuracy of the VQA system. The models under scrutiny in this analysis encompass various methods of feature extraction that help to improve the model and quality of the healthcare industry. The researcher will find the proper methodology for different datasets. To gauge the efficacy of each model in delivering the desired outcomes, an array of metrics will be employed, including classification measurement accuracy, F-classification, C-true positive rate (CTPR), C-precision, C-recall, C-sensitivity, and C-false negative rate (FNR). These metrics are designed to enhance the accuracy of any dataset and optimize the performance of both visual and textual components to ensure accurate responses to the posed queries.

Keywords- Medical Images, VQA, Visual and Textual Feature Extraction methods, Classification model.

I. INTRODUCTION

The field of medical science is experiencing rapid expansion, with a multitude of methods and strategies aimed at enhancing the welfare of patients, researchers, and clinicians alike. In recent years, the convergence of medical and computer science research has given rise to intelligent systems designed to facilitate medical decision-making. Diverse software solutions have been introduced by various providers to aid clinicians, patients, and healthcare practitioners. Researchers are enthusiastically embracing technology to pioneer novel approaches with the potential to benefit society.

Patients often grapple with comprehending the intricacies of their physical and medical conditions. In this context, the Visual Question Answering System has emerged as a prominent and invaluable research tool. This system finds its primary application in the realm of developing solutions capable of responding to queries based on visual imagery. The adoption of this technique has significantly bolstered decision-making processes across various domains and advanced technological applications.

The contemporary medical landscape is marked by swift expansion, encompassing the comprehensive scanning of the

human body through cutting-edge methodologies. While these scan datasets are predominantly in image format, manually deciphering the underlying textual context to address patient inquiries can be a daunting task. Within this context, our research focuses on medical image analysis, particularly within the domain of human skeletal imagery, leveraging an array of datasets available in the medical field.

The principal objective of this study is to identify and harness diverse datasets that facilitate the application of the Visual Question Answering (VQA) system. Additionally, this research seeks to assist medical professionals in making informed decisions while also providing valuable insights to researchers concerning system performance, thereby facilitating improvements through the development of new models catering to both visual and textual information.

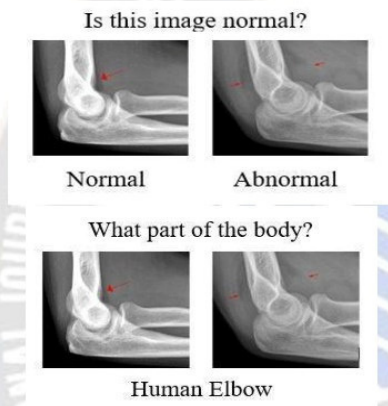


Figure 1. Actual Question Answer from Skeletal Image

In the realm of healthcare, there have been numerous advancements aimed at enhancing accessibility and facilitating medical assistance. Visual Question Answering (VQA) represents a unique approach that offers substantial benefits to a diverse range of patients. This method empowers individuals with the ability to conduct independent research on their medical conditions, reducing their dependency on healthcare professionals.

Over the years, computer technology has become increasingly prevalent within the healthcare sector, playing pivotal roles in various medical services. With the incorporation of VQA, patient-assistance systems are poised to significantly enhance the clarity and comprehension of diverse radiological image types.

Our proposed system is tailored to the specific domain of Skeletal Scintigraphy, encompassing a wide array of topics such as bone marrow, bone cancer, bone density, infections, osteonecrosis, osteoporosis, and more. This system not only assists patients in understanding these complex medical issues but also includes a multilingual feature to accommodate

individuals with limited literacy skills, ensuring inclusivity and accessibility for a diverse patient population.

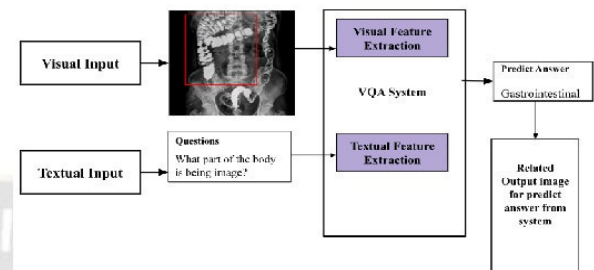


Figure 2. Visual Question Answer system for visual and textual input with Predictive answer

II. PROBLEM STATEMENT FOR THE VQA SYSTEM

One of the difficult tasks in the medical industry is deriving useful information from medical imaging. The fundamental technology of question-answer systems is the extraction of precise user responses. Similar to the quickly expanding medical domain system, the input data extraction process needs to produce an effective and user-satisfying result. The most important component in classifying texts and images is feature extraction, which necessitates a deep understanding of the geometry and forms of real-world objects. Several classification methods entail performing data preprocessing operations, including normalization, identifying the classes, and extracting important features from the data cubes. In addition to making it easier for users to get images of any kind, the objective of solving VQA-related issues is to improve the description of the images and the accuracy of the related images by providing answers to the questions. For ease of understanding and traceability, the process of responding to the inquiry ought to be more descriptive.


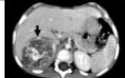
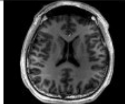
This technique aids in determining the kinds of images that every scanner captures. It is easiest to write below the answers in the visual technique for the corresponding questions when the visualization technique projects the answers as a baseline and displays the relevant region with numerous colors. This type of method yields the highest precision. In order to transform and construct a model utilizing classification, this suggested framework focuses on radiological imaging for bone scintigraphy. According to these ideas, the most useful medical methods for providing visual answers are those that help physicians with clinical analysis and diagnosis. Additionally, this will support hospital services in growing the medical field. Applications for classification techniques can be found in a variety of disciplines, such as traffic identification, medicine, and security. The textual and visual features can be extracted using the feature extraction model. For providing visual answers to questions about radiological imaging, it is the most effective approach. In this paper, we mentioned various datasets, various feature extraction methods, and their accuracy in the healthcare domain.


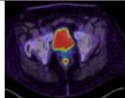
III. RELATED WORK

There doesn't seem to be much research on VQA innovation that focuses on domains of professional healthcare yet. The visual mechanism of radiology images is complex, and the effect on visual perception is modest. Because people are anatomically similar, most radiological scans of the same body location in various people are rather comparable. Although the images seem to be the same, our inspection will reveal different problems. Furthermore, as Figure 1 illustrates, there are multiple questions that have been trained to model different responses. In [2], current datasets, sources, data quantities, and profession features were highlighted. Approaches were reviewed, suggestions were summarized, and future enhancements were planned. It encourages researchers working in this field. Figure 2 illustrates the structure of proposed design which has visual and textual inputs where medical or skeletal radiology images are considered as visual input and the questions related to input image are considered as textual input. The VQA system will extract the features of the input image using the Faster RCNN method and for text we preprocess the data, then extract the key features and finally integrate with classification methods for predicated output.

Radiology is a branch of medicine that uses imaging methods to identify, diagnose, and treat illnesses [8]. Two subspecialties of radiology are diagnostic radiology and interventional radiology [1]. Radiologists can assess internal body components using diagnostic radiography to seek out health problems, assess the source of symptoms, and track the body's response to treatment. The radiological modalities that are most frequently utilized are positron emission tomography (PET), magnetic resonance tomography, computed axial tomography, plain radiographic images, and ultrasound imaging [9]. It is helpful to visualize a variety of illnesses, including heart disease, colon cancer, and breast cancer. One of the most commonly used kinds of diagnostic radiology scans is CT (computerized tomography), also referred to as CAT (computerized axial tomography). Table 1 enumerates the diverse categories of radiological images alongside the corresponding medical terminology names for our system.

TABLE 1 MEDICAL TERMINOLOGY FOR THE VQA SYSTEM

TYPES OF RADIOLOGY IMAGES	NAME OF IMAGE	IMAGES
X-ray Image	cervical spine	
CT Scan Image	Abdominal	
MRI Scan Image	Human cerebrum	

Ultrasound Image	Fetal	
PET Image	Sarcoma	

A. Challenges in Healthcare Datasets

Large-scale medical dataset preparation will require a great deal of work, and it should be done with due consideration for clinicians or physicians. Developing a medical VQA dataset is a highly challenging task. When creating a dataset, it is important to include photos from different radiology specialties, classify clinical questions for each image, have a solid understanding of medical terminology, and create precise responses for each question. We must lower the noise level of both the categorized question and answer because the noise level of the constructed dataset will be high. The dataset also includes a large number of photos with unclear pixels, objects, and other image errors. So it will be of absolutely no help for medical treatment, and it also includes questions that patients will find incomprehensible. Every image and response should follow the correct structure so that medical professionals can understand it. Another problem in the medical arena is scaling up the method to all unlabeled photos in the healthcare dataset.

B. Challenges in Feature Extraction Model

The existing Visual Question Answering (VQA) models employ Convolutional Neural Networks (CNN) to extract local regional vectors for specific areas within images. Long Short-Term Memory (LSTM) models are utilized to encode the feature vectors corresponding to the questions posed. While these models perform admirably in generating answers, they encounter limitations in scenarios where the response involves two adjacent local regions in the image and the question is structured as a complex sentence. It's worth noting that these models do not factor in the position and orientation of objects in their predictions.

Additionally, it's important to acknowledge that convolution operations are computationally more intensive and slower compared to max-pooling operations, both during forward and backward passes. Consequently, when dealing with deep networks, each training iteration naturally demands a substantially longer duration.

CNN-based algorithms necessitate extensive datasets to produce meaningful results, a limitation that can be challenging when dealing with scenarios involving a limited number of training instances. This is particularly significant considering the considerable resources, including time and expertise, required to compile and accurately categorize a comprehensive collection of images. In such cases, techniques like "data augmentation" and "transfer learning" are employed to address these limitations. Effective categorization heavily depends on the correct selection of image properties, as even the most

advanced machine-learning classifiers may perform poorly if these attributes are not appropriately chosen.

In addressing the vanishing gradient problem, Long Short-Term Memory (LSTM) models represent a noteworthy improvement over traditional Recurrent Neural Networks (RNNs). They expand the memory of RNNs to capture and retain long-term input dependencies. The "gated" cell within LSTM models empowers them to read, write, and erase information from memory, making informed decisions about which information to preserve or disregard.

The BiLSTM-CNN model employs Bidirectional LSTMs to encode both past and future contexts at each time step, following the CNN's encoding of each word. While this is beneficial for tasks like machine translation and sentence classification, it poses limitations for sequence-labeling tasks such as Named Entity Recognition (NER), as each token utilizes its own midway hidden states, unable to bridge past and future context effectively.

This research encompassed diverse datasets, various image feature extraction models, and textual feature extraction models, with summarized results presented in the following table.

TABLE 2 INSIGHTS FROM THE LITERATURE SURVEY WITH VARIOUS DATASETS, FEATURE EXTRACTION AND ITS ACCURACY

MODEL	DATASETS	IMAGE FEATURE EXTRACTION	TEXT FE	CLASSIFICATION	CATEGORIES OF QUESTION
Visual-Language Model	VQA-RAD and PathVQA	ViT32 Model	BERT	Contrastive language-image pre-training (CLIP) model	Open-ended, Closed-ended
BPMVQA	VQA-Med 2018, ImageCLEF 2019, VQA-RAD	CNN model to extract the spatial features	PubMed	Self-attention module and a feed forward neural network (FFN)	What, where, Yes/No
MedFusionNet	ImageCLEF 2019	CNN models	BERT	MFB	modality

					Plane Organ
Adversarial VQA benchmark	Human-And-Model-in-the-Loop Enabled Training (HAMLET)	Adversarial VQA	SOTA models	-	Counting, OCR, Reasoning, Visual concept recognition

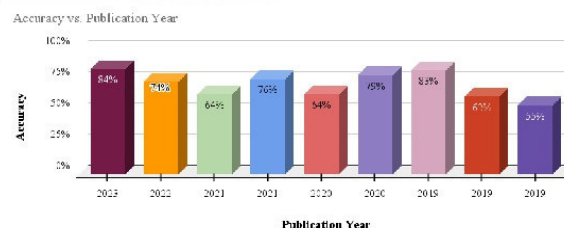


Figure 3. Existing Accuracy based on Publication Year

IV. METHODOLOGY

The system records any kind of radiological image; there are no restrictions on the kind of image that can be chosen when responding visually to questions related to bone scintigraphy. Diagnostic and interventional radiology images are two distinct categories. It is an imaging technology that helps in illness diagnosis and treatment. The system is designed to examine the skeletal scan, which is the equivalent of the bone scan aids in the detection of numerous conditions, including bone joint disorders, insufficiency fractures, shattered bones, and bone cancer. This provides an answer to the issue for every kind of bone in the human body, including long, short, and irregular bones, from the head to the foot. To make the process of asking and answering questions easier to comprehend, there should be more description in the process. This makes it easier for all patients and physicians to view the images clearly and eliminates the majority of doubts with a thorough description. This system serves the purpose of categorizing images produced by medical imaging tools, distinguishing between images from diagnostic radiology and interventional radiology. It enables users to pose questions related to these regions. As described in Figure 1, it provides answers based on user-generated questions and retrieves pertinent images in response. This feature greatly enhances user understanding and facilitates follow-up care. The referenced images inferred from the answers are consistent with the image-based responses.

A. Use case related to Proposed System

The fetal skull organ, as illustrated in Figure 3, is the subject of discussion. In this context, the system enables users to input an image featuring two distinct perspectives: the superior view and

the lateral view. The system furnishes the results in Table 3, presenting details such as the questions posed, the types of questions (both objective and subjective), the identified organ, and the image type, as detailed in Table 2. Upon selecting any of the organ names listed in the table, users will be directed to the section of the displayed image marked by a red line at its base. Furthermore, the relevant organ names associated with the displayed image segment will be presented to users on the same page, in line with the depiction in Figure 3. Users will also have the option to access additional information regarding the image via the related image selection, conveniently available on the same screen. This approach aims to offer users a seamless and informative experience.

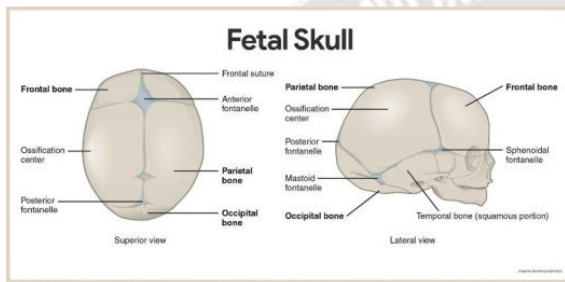


Figure 4. Fetal Skull in Superior view and Lateral view

The collection contains about 3000 radiology and skeletal images from the medical domain. The training, testing, and cross validation ratios are 70:30. 70% of images from the preprocessed dataset will be used for training, while the remaining 30% will be used for testing and cross validation. In the same ratio, questions and answers will be trained from preprocessed images.

TABLE 3 FORMULATED SINGLE IMAGE WITH VARIOUS QUESTIONS

Questions	Objective Answers	Subjective Answers	Organ	Image Type
What does the CT scan show?	left atrium	A large filling defect in the left atrium.	Fetal Skull	Diagnostic
Where is the anterior fontanel?	Top		Fetal Skull	Diagnostic

Is it normal?	Yes		Fetal Skull	Diagnostic
---------------	-----	--	-------------	------------

1) Visual and Textual Feature Extraction

Most cutting-edge medical VQA systems rely on deep learning methods like attention mechanisms and recurrent neural networks (RNNs) [12] for text embedding and feature extraction, and convolutional neural networks (CNNs) for visual feature extraction. Deep learning transformers have been developed and successfully used for the medical VQA requirement. Transformers, for example, were originally applied to NLP applications like speech recognition [14] and machine translation [13]. The self-attention mechanism is the only source of dependency for its encoder-decoder design. Transformers show promise in learning relationships among sequence elements, in contrast to RNNs, which process sequence items recursively and only consider immediate context. Transformer designs that focus on entire sequences have the potential to learn long-range correlations. Specifically, the most commonly used model for textual information encoding is the bidirectional encoder representation from transformers (BERT) [15]. Using large-scale unsupervised corpora and a bidirectional attention mechanism, the language model BERT generates a context-sensitive representation for every word in a sentence. R-CNN's limitations were addressed with the introduction of Fast R-CNN. To create a convolutional feature map in this case, we simply send the input to CNN. From there, we identify the region proposals and use the ROI pooling layer to warp them into squares. The size can be changed, and it can feed into layers that are completely connected. It feeds none of the 2000 areas. According to the image, it immediately created the feature map. Compared to RCC, it is much faster for testing and training.

A faster R-CNN, also known as Fast R-CNN, is employed to identify region suggestions for selective search. It increases training and testing speed. The time it takes to get the output is decreased. To find the region proposals, a convolutional feature map performs better than a selective search technique.

The deep belief network training algorithm DBN can be used to initialize the network with random weights. Next, unsupervised learning can be used to train each layer of the network, starting from the first layer and continuing through the last layer. Finally, supervised learning and backpropagation can be used to fine-tune the entire network. This process must be repeated until the network has converged.

BiLSTM, or Bidirectional Long Short-Term Memory, comprises two separate LSTM neural networks, each with its own unique set of weights and bias factors. The outputs from the hidden layers of the forward and backward networks are combined through concatenation to form the feature vector that is subsequently extracted. In a study conducted by Linqin Cai, Sitong Zhou, Xun Yan, and Rongdi Yuan in 2019, they extensively discuss the operation of the Stacked Bidirectional Long Short-Term Memory Neural Network (SBiLSTMNN). They also delve into the coattention mechanism for question

representation and the attentive attention mechanism for answer representation. This comprehensive approach aims to provide a deep understanding of the SBiLSTMNN and its associated mechanisms.

2) Analysis of Experimental Results

Various datasets were analyzed from various research papers, which are mentioned in Table 2. Different categories of data that were used, its question answer type, and images, as shown in Fig. 4. It describes the total amount of data that leads to the ratio needed to train the model.

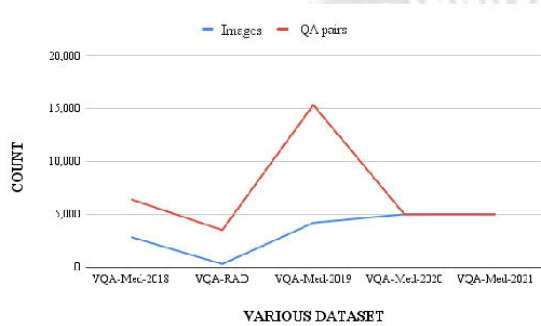


Figure 5. Types of datasets with total count of images and text mentioned in Table 2

Figure 5 depicts the total number of questions and images available to trained models. Each image has numerous questions, each in its own category.

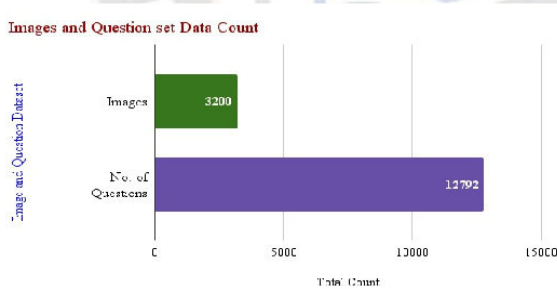


Figure 6. Visual and Textual Datasets

The many sorts of questions are depicted in the image below; each includes over 3000 question and answer sets to train the model, which is sufficient to develop the system. This helps to categorize the question and makes it easier to find related responses to the user's question. This type of question and answer was employed in the majority of previous models. Figure 6 illustrates the cumulative count of questions within each dataset category, each encompassing more than 3000 questions.

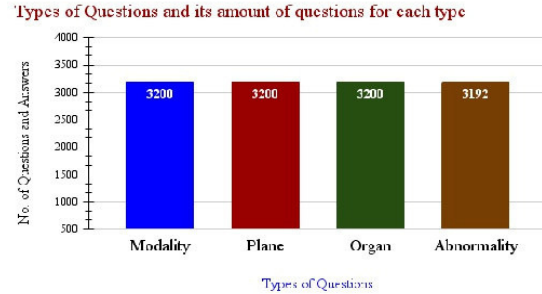


Figure 7. Category of Questions from dataset with a Total Count for each type

B. Performance Metrics

The assessment of each existing model's performance involves the consideration of several key metrics, such as mean absolute error, mean squared error, root mean squared error, false measure, and precision.

Mean squared error (MSE) is a crucial metric that enables us to determine the average of the squared differences between the ground truth value (Y_j) and the predicted regression value (Y'). N represents the number of data points, as per equation (1).

In contrast, the mean absolute error (MAE) calculates the average of the differences between the ground truth and the predicted values, providing insights into the extent of deviations between forecasts and actual outcomes. It's worth noting that MAE employs the absolute value of the residual, making it direction-agnostic, meaning it doesn't discern whether under- or over-prediction has occurred. As outlined in equation (2), MAE is particularly robust against the influence of outliers.

This formal evaluation methodology aims to rigorously assess and compare the performance of these models in a quantifiable manner.

$$MAE = \frac{1}{N} \sum_{j=1}^N (Y_j - Y'_j)^2 \quad (1)$$

$$MSE = \frac{1}{N} \sum_{j=1}^N |Y_j - Y'_j| \quad (2)$$

The root mean squared error (RMSE) plays a significant role in the assessment of model performance. It calculates the average of the squared differences between the target value and the value predicted by the regression model. RMSE is particularly valuable because it rectifies a potential limitation of MSE, which excessively penalizes smaller errors by taking the square root of the result.

This square root transformation ensures that the scale of error interpretation aligns with that of the random variable, simplifying the process of understanding and analyzing errors. Essentially, RMSE normalizes the variables, reducing the potential impact of outliers on the overall analysis. This normalization is exemplified in equation (3).

In a formal evaluation context, RMSE provides an effective means of assessing the models, taking into account the scale of errors, and facilitating their meaningful interpretation.

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^N (Y_j - Y'_j)^2} \quad (3)$$

The Fmeasure range of feasible feature extraction approaches for both visual and textual datasets is depicted in Figure 7. For our datasets, the basic CNN has a lower level of Fmeasure than RNN and DBN.

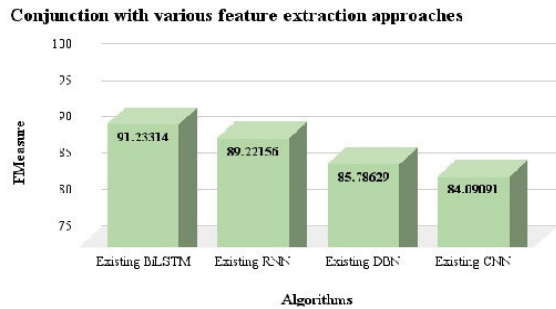


Figure 8. Conjunction with various feature extraction models

C. Significance of the study

The exploratory outcome of the content extraction study, as shown in Figure 8, uses the removal to calculate the degree of coordination between the inquiry vector and the reaction vector. Manjunath Jogin and Mohana, May 2018, investigation study for execution of various categorization computations in Table 3. Consider the present models that have low accuracy for image highlight extraction and question reply feature extraction in Jinesh Melvin Y I, Sushopti Gawade, and Hemant Palivela, May 2021. The goal of the same paper was to describe the Visual Address Replying Framework for Radiology Images from Human Skeletal.

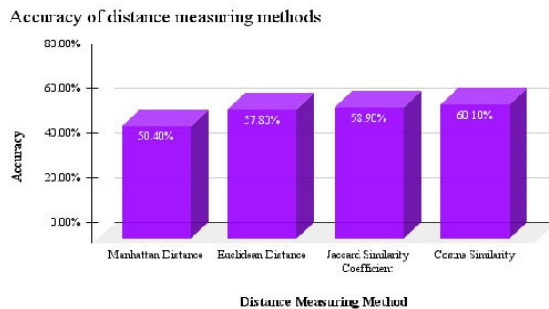


Figure 9. Accuracy with its distance measuring method for classifying various datasets

V. CONCLUSION

A comparative analysis of diverse feature extraction methodologies was conducted, employing a range of distinct

datasets. This analytical approach serves to provide valuable insights for researchers striving to enhance the healthcare system's diagnostic capabilities for patient health assessment. Datasets were meticulously collected from a multitude of sources, facilitating a comprehensive evaluation of the existing methodologies within question-answering systems. The intended outcome of this endeavor is to contribute to the advancement of healthcare, ultimately enhancing the efficiency and effectiveness of patient outcomes. The future adoption and utilization of medical Visual Question Answering (VQA) systems will be contingent upon several pivotal factors. These factors encompass the abundance and caliber of medical VQA datasets, the development and evaluation of medical VQA models, as well as the seamless integration and practical deployment of medical VQA systems within clinical contexts. A critical imperative involves the generation of expansive, comprehensive, and heterogeneous medical VQA datasets that encompass a diverse spectrum of modalities, medical conditions, question types, and corresponding responses.

VI. DECLARATIONS

A. Funding

The authors specifically state that they received no financial aid, grants, or other forms of assistance to facilitate their research. This declaration emphasizes the research's independence and lack of outside influences on its findings.

B. Statement on Conflicts of Interest

This work's authors have reported no conflicts of interest connected to the subject matter.

C. Ethics Declaration

The author explicitly declares a lack of awareness regarding any personal or professional conflicts that might have influenced the research presented in this study. This statement underscores the commitment to maintaining impartiality and objectivity in the research.

D. Code and Data Availability Statement

We used data from a variety of publicly available sources for the research, such as medical visual question answers from CLEF. This allows us to evaluate a variety of existing models and create a new framework for our system. The custom code is used to develop the application, which is used by us. The code for this project is confidential.

AUTHORS CONTRIBUTION STATEMENT

Jinesh Melvin Y. I. is the corresponding author for the said manuscript. Jinesh Melvin Y.I. and Sushopti Gawade conceived of the presented idea. Jinesh Melvin Y. I. developed the theory and performed the computations. Sushopti Gawade and Mukesh Shrimali verified the analytical methods, encouraged Jinesh Melvin Y I to investigate the proposed work, and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

Jinesh Melvin Y I developed the theoretical formalism, performed the analytic calculations, and performed the numerical simulations. Both Jinesh Melvin Y I, Sushopti

Gawade, and Mukesh Shrimali contributed to the final version of the manuscript. Sushopti Gawade and Mukesh Shrimali supervised the project.

REFERENCES

- [1] Y. I. Jinesh Melvin, Sushopti Gawade, Hemant Palivela, "Feature Extraction from Radiology Images for Visual Question Answering System Using CNN and BiLSTM Model", *Recent Innovations in Computing*, vol.832, pp.317, 2022.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [2] Yakoub Bazil, Mohamad Mahmoud Al Rahhal 2, Laila Bashmal 1 and Mansour Zuair 1 "Vision–Language Model for Visual Question Answering in Medical Imagery", *Bioengineering* 2023.
- [3] Stefania Barburiceanu, Serban Meza, Bogdan Orza, Raul Malutan, Romulus Terebes."Convolutional Neural Networks for Texture Feature Extraction. Applications to Leaf Disease Classification in Precision Agriculture", *IEEE Access*, 2021.
- [4] Y. Lu and S. Young, "A survey of public datasets for computer vision tasks in precision agriculture", *Comput. Electron. Agricult.*, vol. 178, Nov. 2020.
- [5] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning", *J. Big Data*, vol. 6, no. 1, pp. 60, 2019.
- [6] N. Ganatra and A. Patel, "A survey on disease detection and classification of agriculture products using image processing and machine learning", *Int. J. Comput. Appl.*, vol. 180, no. 13, pp. 7-12, Jan. 2018.
- [7] M. D. Zeiler, R. Fergus, "Visualizing and understanding convolutional networks", *ECCV*, 2014.
- [8] Herring W, *Learning radiology: Recognizing the basics*. Elsevier Health Sciences, 2015.
- [9] Novelline RA and Squire LF, *Squire's fundamentals of radiology*. La Editorial, UPR, 2004.
- [10] N. Ganatra and A. Patel, "A survey on disease detection and classification of agriculture products using image processing and machine learning", *Int. J. Comput. Appl.*, vol. 180, no. 13, pp. 7-12, Jan. 2018.
- [11] Sima Siami-Namini, Neda Tavakoli, Akbar Siami Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series ", *IEEE International Conference on Big Data (Big Data)* 2019.
- [12] Mikolov, T.; Karafiat, M.; Burget, L.; Cernocky, J.; Khudanpur, S. Recurrent Neural Network Based Language Model. *Interspeech* 2010, 2, 1045–1048.
- [13] Wang, Q.; Li, B.; Xiao, T.; Zhu, J.; Li, C.; Wong, D.F.; Chao, L.S. Learning Deep Transformer Models for Machine Translation. *arXiv* 2019, arXiv:1906.01787.
- [14] Chen, N.; Watanabe, S.; Villalba, J.A.; Zelasko, P.; Dehak, N. Non-Autoregressive Transformer for Speech Recognition. *IEEE Signal Process. Lett.* 2021, 28, 121–125.
- [15] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* 2019, arXiv:1810.04805.
- [16] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu, "ImageCLEF 2019 Visual Question Answering in the Medical Domain," *Zhejiang University, Hangzhou, China*, Sep 2019.
- [17] Lubna A, Saidalavi Kalady, Lijiya A., "MoBVQA: A Modality based Medical Image Visual Question Answering System", 978-1-7281-1895-6/19/\$31.00 c 2019 IEEE, 2019 IEEE Region 10 Conference (TENCON 2019).
- [18] Asma Ben Abacha, Soumya Gayen, Jason J Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman, "NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain", *CEUR-WS.org/Vol 2125/paper_165.pdf*, Conference Paper · October 2018
- [19] Manjunath Jogin, Mohana, Madhulika M S, Divya G D, Meghana R K, Apoorva S, "Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning", 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT-2018), MAY 18th & 19th 2018.
- [20] Zhou Yu, Jun Yu, Jianping Fan, Dacheng Tao, "Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering", *arXiv:1708.01471v1 [cs.CV]* 4 Aug 2017

CERTIFICATES



CERTIFICATE OF PRESENTATION

This certificate is awarded to

Mr. Jinesh Melvin Y I

for successfully presenting a paper at the


International Conference on Artificial Intelligence and Smart Systems (ICAIS 2021)
organised by JCT College of Engineering and Technology, Coimbatore, India
on 25-27, March 2021.

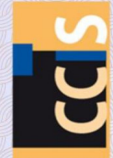
Paper Title: Visual Question Answering using Data Mining Techniques for Skeletal Scintigraphy in medical domain - VQADMSS

Author/s: Mr. Jinesh Melvin Y I, Dr. Sushopti Gawade, Dr. Hemant Palivela


Session Chair


Conference Chair
Dr. K. Geetha


Principal
Dr. V. J. Arulkarthick



**1st Springer CCIS International Conference on
Role of AI in Bio-Medical Translations' Research for the Health Care Industry**

Organized by

G H RAISONI COLLEGE OF ENGINEERING, NAGPUR

Certificate

to

Jinesh Melvin Y I, Sushopti Gawade

**Visual Question Answering System for Skeletal Image based on
feature Extraction using Faster RCNN AND Kai-Bi-LSTM Techniques.**

Paper Title:

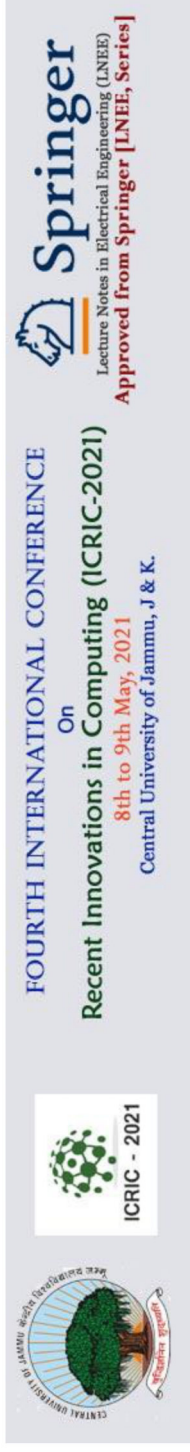
has presented in 1st Springer CCIS International Conference on **"Role of AI in Bio-Medical Translations' Research for the Health Care Industry"**(AIBTR-23) held on 23rd September, 2023 organized by Center of Excellence Biomedical Engineering & Technology incubation Center (BETiC-GHRCE) and Artificial Intelligence & Machine Learning (AIML) of G H Raisoni College of Engineering, Nagpur (India).

Dr. P. Sivaram
General Chair

Dr. Vikas Bora
General Chair

Dr. Sachin Untawale
Honorary Chair & Director, GHRCE





FOURTH INTERNATIONAL CONFERENCE

On

Recent Innovations in Computing (ICRIC-2021)

8th to 9th May, 2021

Central University of Jammu, J & K.

Fourth International Conference on Recent Innovation in Computing

ICRIC-2021

May 08-09, 2021

Organized by

Department of Computer Science & Information Technology

CENTRAL UNIVERSITY OF JAMMU, J&K, INDIA

Certificate of Presentation/Participation

This is to certify that Mr./Ms. Jinesh Melvin Y I of Pillai College of Engineering participated/presented a paper entitled **Feature Extraction from Radiology images for Visual Question Answering system using CNN and BiLSTM model.** during 4th International Conference on Recent Innovation in Computing organised by Central University of Jammu, Jammu, J & K, India.

A handwritten signature in blue ink, appearing to read 'Dr. Yashwant Singh'.

Dr. Yashwant Singh
General Chair



दिनांक/Dated: 07/02/2024

SUSHOPTI GAWADE, F8/0-1, INDIANCOMETAX COLONY,
SECTOR 21 22, CBD BELAPUR, NAVI MUMBAI
MAHARASHTRA-400614
INDIAN
JINSH MELVIN Y.I, SECTOR 5A, ROOM NO 1003, PLOT NO
59, KARANJADE, MAHARASHTRA-410206
INDIAN

COMPUTER SOFTWARE WORK DEVELOPMENT OF A MEDICAL VISUAL QUESTION ANSWERING MEDVQA SYSTEM THROUGH THE INTEGRATION OF ADVANCED FEATURE EXTRACTION METHODS ON RADIOLOGICAL IMAGES.

DEVELOPMENT OF A MEDICAL VISUAL QUESTION ANSWERING MEDVQA SYSTEM THROUGH THE INTEGRATION OF ADVANCED FEATURE EXTRACTION METHODS ON RADIOLOGICAL IMAGES.

ENGLISH

SUSHOPTI GAWADE , F8 / 0-1, INDIANCOMETAX COLONY,
SECTOR 21 22, CBD BELAPUR, NAVI MUMBAI
MAHARASHTRA-400614
INDIAN

JINESH MELVIN Y I, SECTOR 5A, ROOM NO 1003, PLOT NO
59, KARANJADE, MAHARASHTRA-410206
INDIAN

UNPUBLISHED

N.A.
का कापीयत, भारत सरकार, इंटेलिक्टुअल प्रपर्टी ऑफिस, गवर्नमेंट ऑफ इंडिया
N.A.
کاپیوارہ ملکیت جو دفتر، دشتیاں جی حکومت، اینڈ پراپرٹی آفیس آف انڈین گورنمنٹ
N.A.
کاپیوارہ ملکیت جو دفتر، دشتیاں جی حکومت، اینڈ پراپرٹی آفیس آف انڈین گورنمنٹ
SUSHOPTI GAWADE, F8 /0-1 INDIANCOMETAX COLONY,
SECTOR 21 22 CBD BELAPUR, NAVI MUMBAI
MAHARASHTRA-400614
INDIAN
JINESH MELVIN Y J., SECTOR 5A, ROOM NO 1003, PLOT NO
59, KARANJADE, MAHARASHTRA-410206
INDIAN

NA

[illegible]

NA

[illegible]

NA

७८६००००६, बौद्धिक सम्पदा कार्यालय, भारत सरकार, मॉपिङ मोंपेरी स्ट्रैट, काठमाडौं

NA

సాంఘిక కార్యాలయము, భారత ప్రభుత్వము, భాగ్య లక్ష్మణం బికానెర్
 గిరి, భారత సర్కార్, Intellectual Property Office, Government of India
 ఈ స్థానము కార్యాచరణ, జాతర వాతావరణ, చిహ్నాల ఆస్తి, టెలిఫోన్, భారత సర్కార్

THANK YOU

Registrar of Copyrights

अप्रोपन की तिथि/Date of Application: 06/01/2024

PLAGIARISM REPORT



VISUAL QUESTION ANSWERING FOR RADIOLOGY IMAGE OF SKELETAL SCINTIGRAPHY IN MEDICAL DOMAIN USING FEATURE EXTRACTION METHOD

by Jinesh Melvin Y I

Submission date: 13-Mar-2024 02:24PM (UTC+0530)

Submission ID: 2319307596

File name: JINESH.pdf (26.12M)

Word count: 36451

Character count: 212210

VISUAL QUESTION ANSWERING FOR RADIOLOGY IMAGE OF SKELETAL SCINTIGRAPHY IN MEDICAL DOMAIN USING FEATURE EXTRACTION METHOD

ORIGINALITY REPORT

9%

SIMILARITY INDEX

6%

INTERNET SOURCES

9%

PUBLICATIONS

2%

STUDENT PAPERS

PRIMARY SOURCES

1	Y. I. Jinesh Melvin, Sushopti Gawade, Mukesh Shrimali. "Chapter 6 Visual Question Answering System for Skeletal Images Based on Feature Extraction Using Faster RCNN and Kai-Bi-LSTM Techniques", Springer Science and Business Media LLC, 2024 Publication	2%
2	www.mdpi.com Internet Source	1%
3	fastercapital.com Internet Source	<1%
4	"ECAI 2020", IOS Press, 2020 Publication	<1%
5	dokumen.pub Internet Source	<1%
6	arxiv.org Internet Source	<1%

7

Internet Source

<1 %

8

Yoshimasa Kawazoe, Kiminori Shimamoto, Ryohei Yamaguchi, Yukako Shintani-Domoto, Hiroshi Uozaki, Masashi Fukayama, Kazuhiko Ohe. "Faster R-CNN-Based Glomerular Detection in Multistained Human Whole Slide Images", Journal of Imaging, 2018

Publication

<1 %

9

[ebin.pub](#)

Internet Source

<1 %

10

[www.javatpoint.com](#)

Internet Source

<1 %

11

Submitted to University of Lancaster

Student Paper

<1 %

12

Submitted to University of West Florida

Student Paper

<1 %

13

Submitted to Liverpool John Moores University

Student Paper

<1 %

14

Lecture Notes in Computer Science, 2015.

Publication

<1 %

15

[ijritcc.org](#)

Internet Source

<1 %

16

[export.arxiv.org](#)

Internet Source

<1 %

17

"Machine Learning and Computational Intelligence Techniques for Data Engineering", Springer Science and Business Media LLC, 2023

Publication

<1 %

18

ceur-ws.org

Internet Source

<1 %

19

www.researchsquare.com

Internet Source

<1 %

20

"Computer Vision and Image Processing", Springer Science and Business Media LLC, 2021

Publication

<1 %

21

Santanu Pattanayak. "Pro Deep Learning with TensorFlow", Springer Science and Business Media LLC, 2017

Publication

<1 %

22

www.gabormelli.com

Internet Source

<1 %

23

Submitted to CSU, San Jose State University

Student Paper

<1 %

24

deepai.org

Internet Source

<1 %

25	"Image and Graphics", Springer Science and Business Media LLC, 2019 Publication	<1 %
26	Submitted to Higher Education Commission Pakistan Student Paper	<1 %
27	Yanzi Zhang. "Relation extraction in Chinese using attention-based bidirectional long short-term memory networks", PeerJ Computer Science, 2023 Publication	<1 %
28	"Computer Vision – ECCV 2016", Springer Nature, 2016 Publication	<1 %
29	opus4.kobv.de Internet Source	<1 %
30	Y I Jinesh Melvin, Sushopti Gawade, Hemant Palivela. "Visual Question Answering using Data Mining Techniques for Skeletal Scintigraphy in medical domain - VQADMSS", 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021 Publication	<1 %
31	towardsdatascience.com Internet Source	<1 %
32	malque.pub Internet Source	<1 %

33	Submitted to Banaras Hindu University Student Paper	<1 %
34	www.cs.cmu.edu Internet Source	<1 %
35	"Computer Vision", Springer Science and Business Media LLC, 2017 Publication	<1 %
36	Lecture Notes in Computer Science, 2016. Publication	<1 %
37	Shengyan Liu, Xuejie Zhang, Xiaobing Zhou, Jian Yang. "BPI-MVQA: a bi-branch model for medical visual question answering", BMC Medical Imaging, 2022 Publication	<1 %
38	www.imarcgroup.com Internet Source	<1 %
39	"Machine Learning and Principles and Practice of Knowledge Discovery in Databases", Springer Science and Business Media LLC, 2021 Publication	<1 %
40	Submitted to University of Newcastle upon Tyne Student Paper	<1 %
41	eprints.whiterose.ac.uk Internet Source	<1 %

42	flore.unifi.it Internet Source	<1 %
43	trepo.tuni.fi Internet Source	<1 %
44	www.politesi.polimi.it Internet Source	<1 %
45	www.researchgate.net Internet Source	<1 %
46	www2.mdpi.com Internet Source	<1 %
47	Submitted to University of Malaya Student Paper	<1 %
48	subscription.packtpub.com Internet Source	<1 %
49	"Artificial Intelligence", Springer Science and Business Media LLC, 2021 Publication	<1 %
50	"Computer Vision – ECCV 2018", Springer Science and Business Media LLC, 2018 Publication	<1 %
51	Submitted to Otto-von-Guericke-Universität Magdeburg Student Paper	<1 %
52	Submitted to University of Stirling Student Paper	<1 %

53	Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, Mansour Zuair. "Vision-Language Model for Visual Question Answering in Medical Imagery", Bioengineering, 2023 Publication	<1 %
54	ijrst.com Internet Source	<1 %
55	rua.ua.es Internet Source	<1 %
56	www.coursehero.com Internet Source	<1 %
57	"Advanced Communication and Intelligent Systems", Springer Science and Business Media LLC, 2023 Publication	<1 %
58	"Experimental IR Meets Multilinguality, Multimodality, and Interaction", Springer Science and Business Media LLC, 2018 Publication	<1 %
59	"Image and Vision Computing", Springer Science and Business Media LLC, 2023 Publication	<1 %
60	"MultiMedia Modeling", Springer Science and Business Media LLC, 2018 Publication	<1 %

61

K.Vishnuvardhan Reddy, B. Kumar, N. Prem Kumar, T. Parasuraman, C.M. Balasubramanian, R. Ramakrishnan. "Chess Match Outcome Prediction via Sequential Data Analysis with Deep Learning", 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), 2023

Publication

<1 %

62

Y. I. Jinesh Melvin, Sushopti Gawade, Hemant Palivela. "Chapter 26 Feature Extraction from Radiology Images for Visual Question Answering System Using CNN and BiLSTM Model", Springer Science and Business Media LLC, 2022

Publication

<1 %

63

openaccess.altinbas.edu.tr

Internet Source

<1 %

64

"Medical Image Computing and Computer Assisted Intervention – MICCAI 2022", Springer Science and Business Media LLC, 2022

Publication

<1 %

65

"Web Information Systems Engineering – WISE 2023", Springer Science and Business Media LLC, 2023

Publication

<1 %

66	Submitted to Berlin School of Business and Innovation Student Paper	<1 %
67	Chaoyang YUE, Wenjin Hu, Fujun Zhang, Xinyue Shi. "Thangka Image Caption Generation Method Combining Multi-scale and Multi-level Aggregation", Research Square Platform LLC, 2024 Publication	<1 %
68	Chien-Hung Chen, Che-Rung Lee, Walter Chen-Hua Lu. "Chapter 2 A Mobile Cloud Framework for Deep Learning and Its Application to Smart Car Camera", Springer Science and Business Media LLC, 2016 Publication	<1 %
69	Sheerin Sitara Noor Mohamed, Kavitha Srinivasan. "A comprehensive interpretation for medical VQA: Datasets, techniques, and challenges", Journal of Intelligent & Fuzzy Systems, 2023 Publication	<1 %
70	Submitted to University of Essex Student Paper	<1 %
71	birmingham.elsevierpure.com Internet Source	<1 %
72	Submitted to ebsu Student Paper	<1 %

73	eprints.qut.edu.au Internet Source	<1 %
74	www.arxiv-vanity.com Internet Source	<1 %
75	www.thieme-connect.com Internet Source	<1 %
76	José Andery Carneiro. "Enhanced tooth segmentation algorithm for panoramic radiographs", Universidade de São Paulo. Agência de Bibliotecas e Coleções Digitais, 2023 Publication	<1 %
77	Submitted to Kingston University Student Paper	<1 %
78	Submitted to Nottingham Trent University Student Paper	<1 %
79	Nowomiejska Katarzyna, Powroźnik Paweł, Paszkowska-Skublewska Maria, Adamczyk Katarzyna et al. "Residual Attention Network for distinction between visible optic disc drusen and healthy optic discs", Optics and Lasers in Engineering, 2024 Publication	<1 %
80	Suheer Al-Hadhrami, Mohamed El Bachir Menai, Saad Al-ahmadi, Ahmed Alnafessah. "A Critical Analysis of Benchmarks, Techniques,	<1 %

and Models in Medical Visual Question Answering", IEEE Access, 2023

Publication

81

Zhiqiang Wan, Haibo He. "AnswerNet: Learning to Answer Questions", IEEE Transactions on Big Data, 2019

Publication

<1 %

82

"Intelligent Systems and Applications", Springer Science and Business Media LLC, 2019

Publication

<1 %

83

"Neural Information Processing", Springer Science and Business Media LLC, 2019

Publication

<1 %

84

Ali Akbar Siddique, Nada Alasbali, Maha Driss, Wadii Boulila, Mohammed S. Alshehri, Jawad Ahmad. "Sustainable collaboration: Federated learning for environmentally conscious forest fire classification in Green Internet of Things (IoT)", Internet of Things, 2024

Publication

<1 %

85

Submitted to Budapest University of Technology and Economics

Student Paper

<1 %

86

medium.com

Internet Source

<1 %

87

www.open-access.bcu.ac.uk

Internet Source

<1 %



www.simplilearn.com
Internet Source

<1 %

Exclude quotes On

Exclude matches < 14 words

Exclude bibliography On