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.

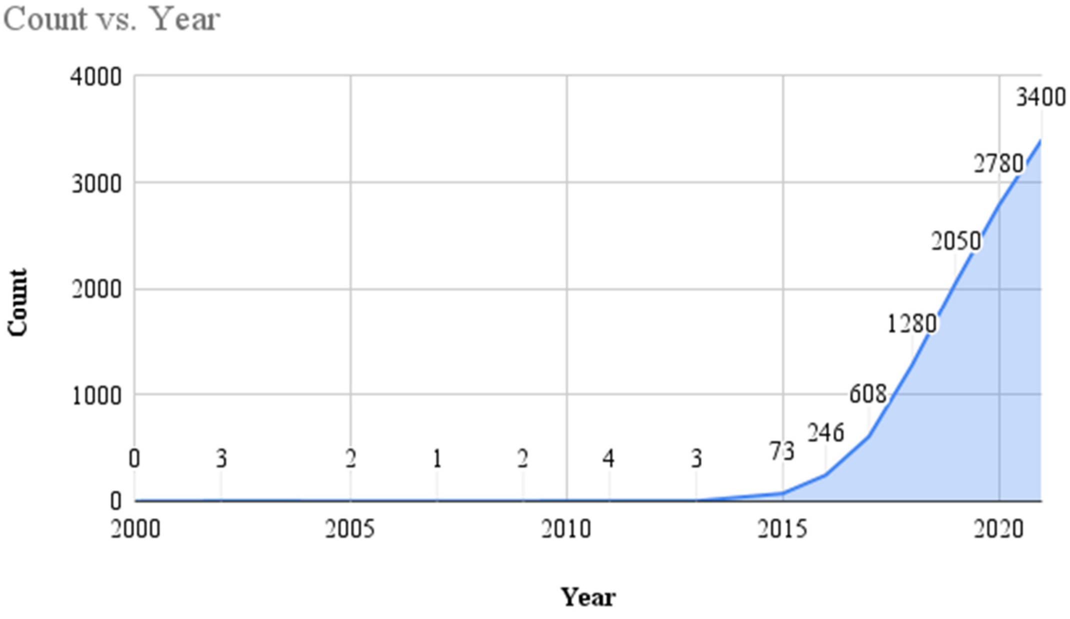

Fig. 2 : Trends in Visual Question Answering Research

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 metic\ulous 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 QuestionAnswering 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 Sliver-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,e 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%

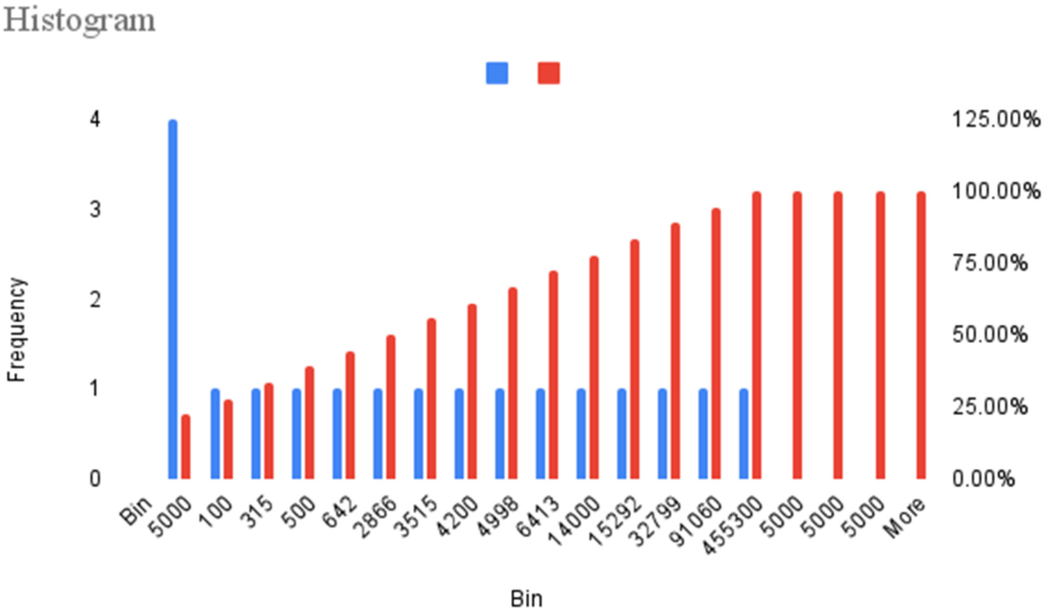

Fig 3 : Histogram of existing collective dataset with its frequency

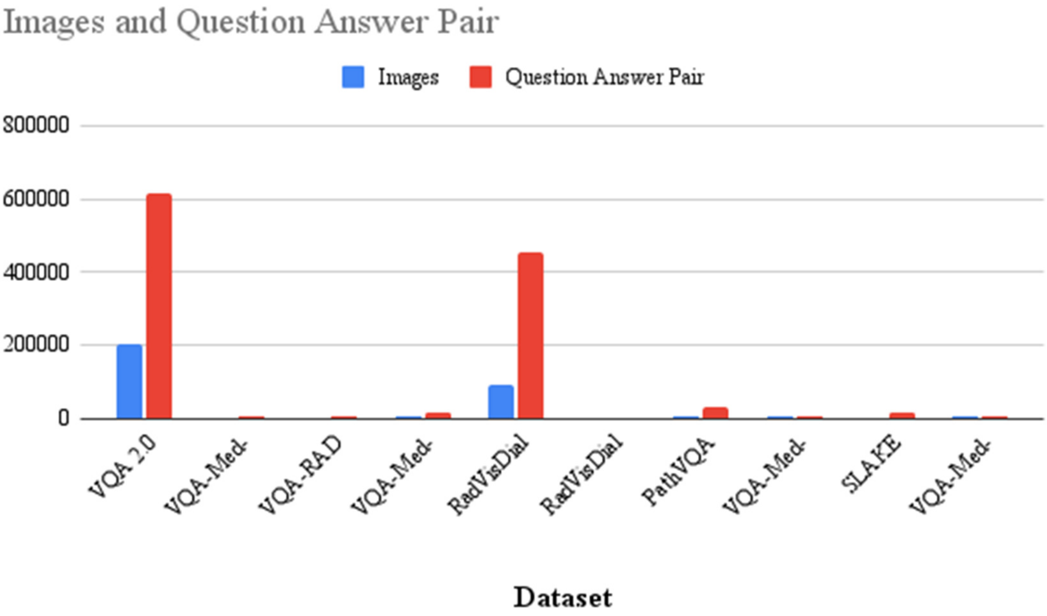

Fig 4 : Total data present in various image and question answer pair dataset

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 imagine 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

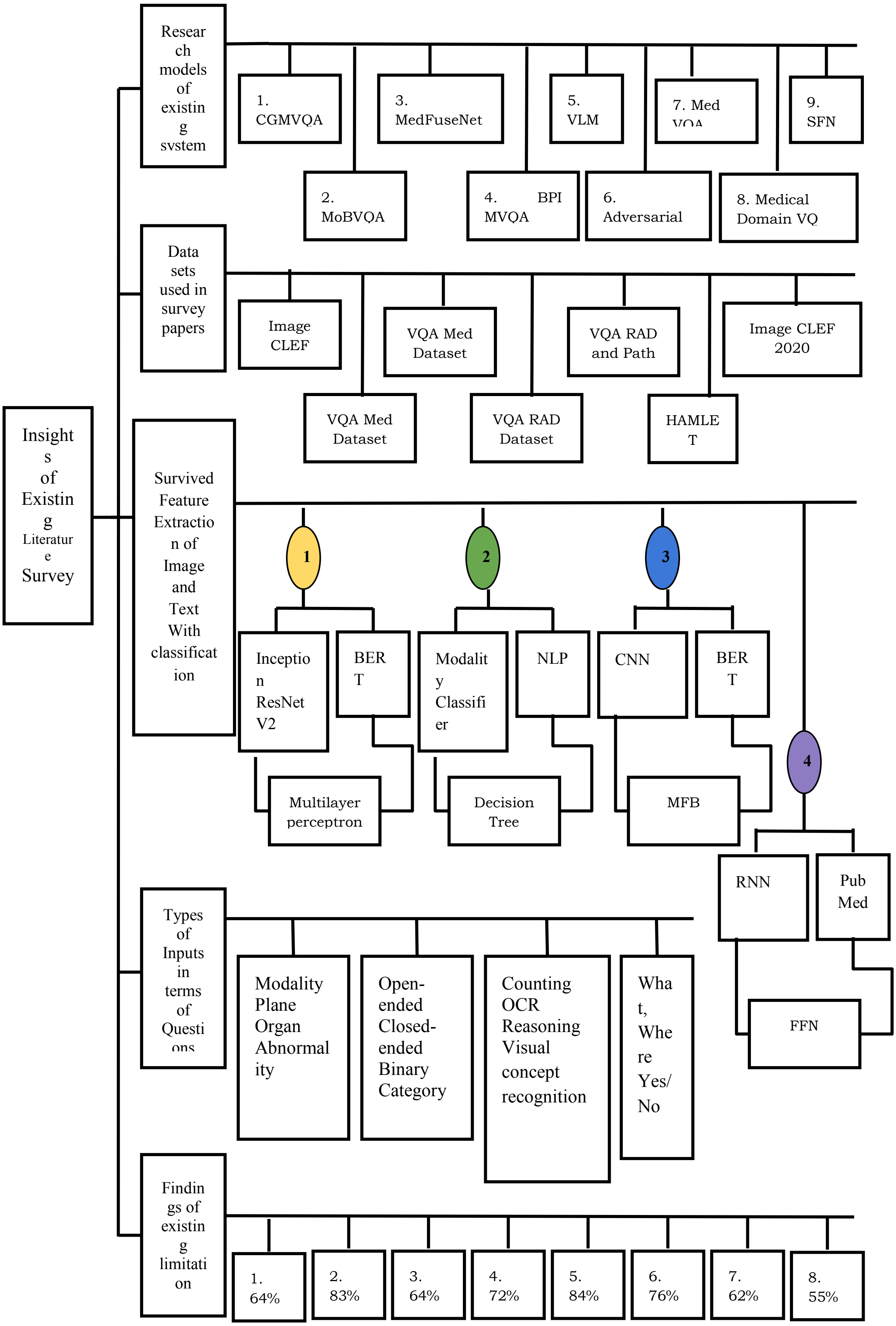

Fig 5 : Insight of Literature Survey

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.