

# 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  
[sushoptiekrishimitra@gmail.com](mailto:sushoptiekrishimitra@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.

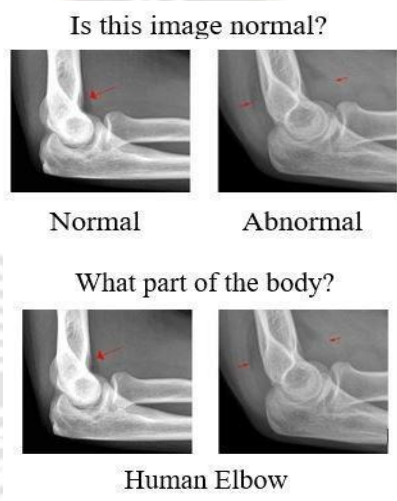


Figure 1. Actual Question Answer from Skeletal Image

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.

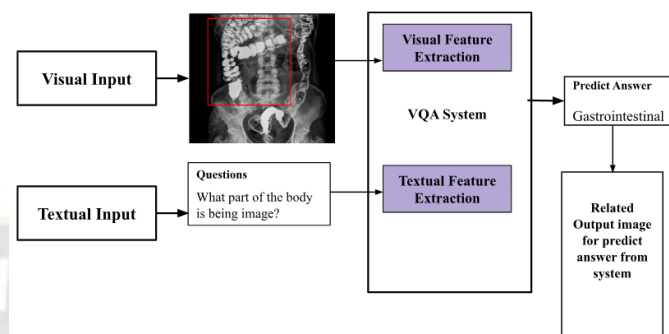


Figure 2. Visual Question Answer system for visual and textual input with Predictive answer

## 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.



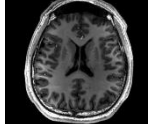
This technique aids in determining the kinds of images that every scanner captures. It is easiest to write below the answers in the visual technique for the corresponding questions when the visualization technique projects the answers as a baseline and displays the relevant region with numerous colors. 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.


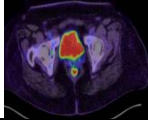
### III. RELATED WORK

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   |
|-----------------------|--|---|---------|---|--------------------------|
| Vision-Language Model | VQA-RAD and PathVQA                    | ViT32 Model                               | BERT    | Contrastive language-image pre-training (CLIP) model          | Open-ended, Closed-ended |
| BPMVQA                | VQA-Med 2018, Image CLEF 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      |
| MedFuset              | Image CLEF 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 |

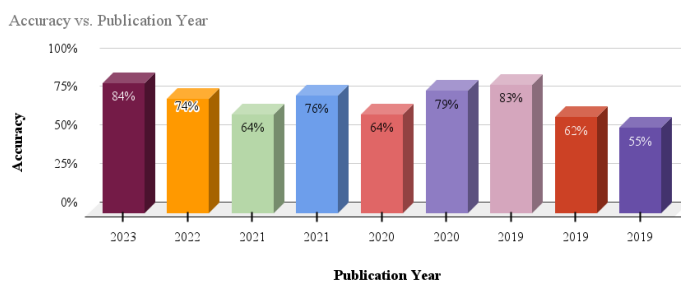


Figure 3. Existing Accuracy based on Publication Year

#### 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.

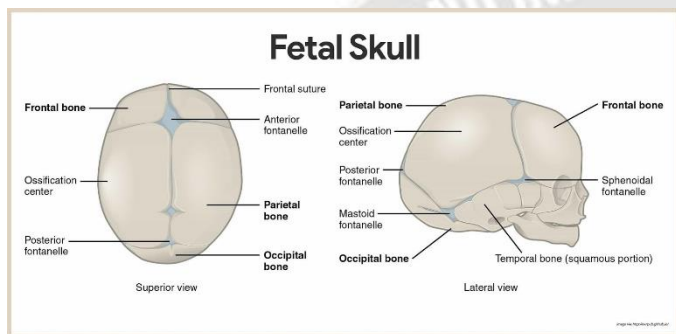


Figure 4. Fetal Skull in Superior view and Lateral view

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.

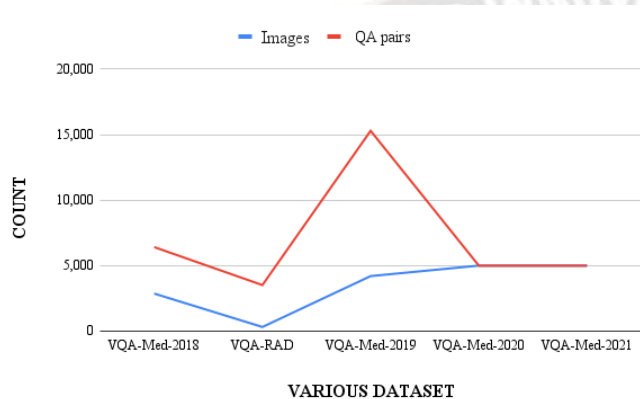


Figure 5. Types of datasets with total count of images and text mentioned in Table 2

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

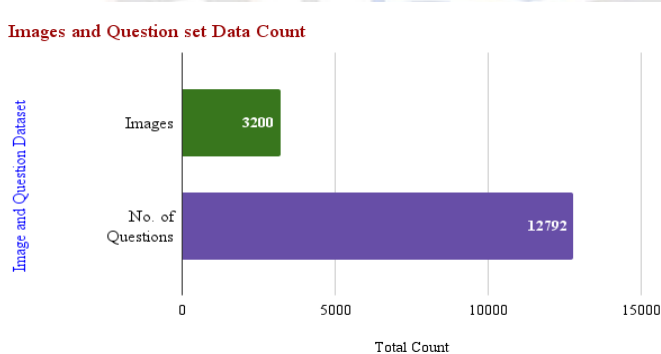


Figure 6. Visual and Textual Datasets

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.

Types of Questions and its amount of questions for each type

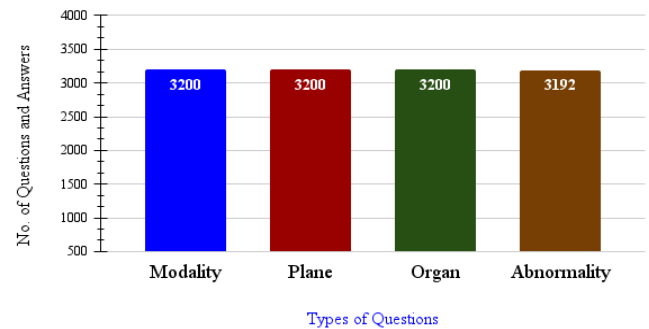


Figure 7. Category of Questions from dataset with a Total Count for each type

## 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.

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

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

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.

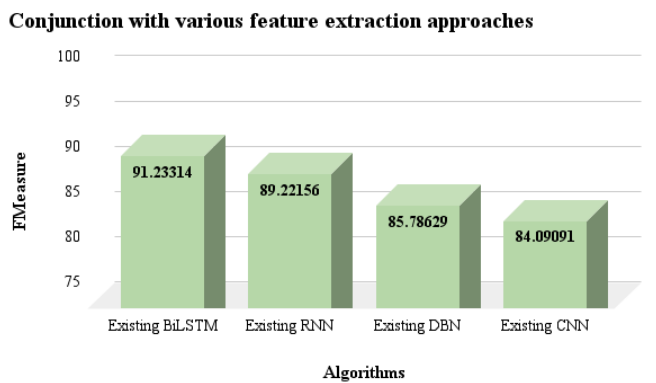


Figure 8. Conjunction with various feature extraction models

### 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.

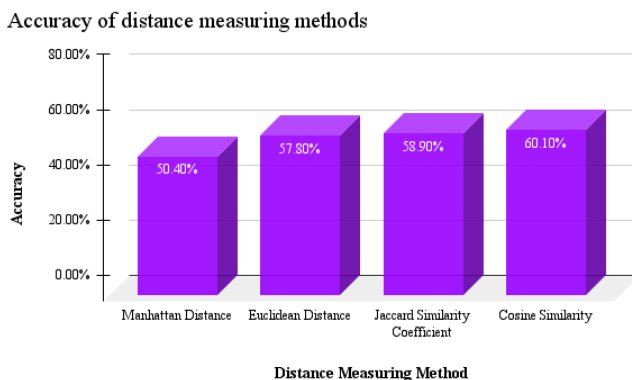


Figure 9. Accuracy with its distance measuring method for classifying various datasets

## 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 Bazi1, 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