Vision-Aided Intelligence with Visual Question Answering for Medical Imaging

Khushi Chalageri[1][0009−0009−1903−8264], Pragatilaxmi Itigowni[2][0009−0005−1273−6669], Disha Kalyanshettar[3][0009−0007−5767−0020], Saakshi Lokhande[4][0009−0006−3394−1618], Channabasappa Muttal[5][0009−0005−7780−3746], and Vaishnavi J Ajjevadeyarmath[6][0009−0003−3874−7329]

[1] School of Computer Science and Engineering, KLE Technological University, Hubli, 580031, India
01fe22bci023@kletech.ac.in

[2] School of Computer Science and Engineering, KLE Technological University, Hubli, 580031, India
01fe22bci013@kletech.ac.in

[3] School of Computer Science and Engineering, KLE Technological University, Hubli, 580031, India
01fe22bci016@kletech.ac.in

[4] School of Computer Science and Engineering, KLE Technological University, Hubli, 580031, India
01fe22bci002@kletech.ac.in

[5] School of Computer Science and Engineering, KLE Technological University, Hubli, 580031, India
channabasappa.muttal@kletech.ac.in

[6] School of Computer Science and Engineering, KLE Technological University, Hubli, 580031, India
ajjevadeyarmathvaishnavi@gmail.com

Abstract. Visual Question Answering(VQA) models have become essential in medical imaging because they enhance patient engagement and aid in clinical decisions. This study introduces a VQA model combining VGG16 for visual feature extraction, Word2Vec for question tokenization, and LSTM for natural language understanding to process fundus images and answer questions regarding Diabetic Macular Edema (DME) grading. The model achieved a training accuracy of 96.88% and a validation accuracy of 87.52%, outperforming ResNet101 in identifying key dataset features. Integrating Word2Vec and LSTM enables the accurate comprehension of complex medical queries. Notably, the model consistently answered related questions without explicit rules. Its simple multimodal fusion approach improves interpretability and computational efficiency without sacrificing performance. Future research could incorporate formal consistency metrics and advanced architectures such as transformers, attention mechanisms, and external knowledge sources to enhance logical coherence, balancing accuracy, and consistency. These results highlight the potential of the proposed VQA model to provide reliable and interpretable outcomes in clinical settings and promote its broader use in medical imaging.

Keywords: Visual Question Answering(VQA), Medical imaging, Natural language questions, Model architectures, Diabetic Macular Edema (DME), Hard exudates, Logical consistency.

1 Introduction

Multimodal learning, which integrates data from various sources, such as visual, textual, and auditory input, has transformed the processing of complex information[1],[2]. By combining diverse data types, these models enable robust, context-aware decision making, making them invaluable in fields such as healthcare, autonomous systems, and human-computer interaction. In VQA these models analyze images and natural language questions and leverage their multimodal capabilities to deliver insightful and contextually relevant answers[3]. This versatility unlocks the transformative potential, particularly in applications such as medical imaging.

To address the specific challenges of medical imaging,we propose a VQA model that integrates VGG16, Word2Vec, and LSTM to assess the severity of diabetic macular edema across various stages, ensuring enhanced accuracy and consistent performance. VQA models integrating image interpretation with natural language understanding are vital in medical imaging to address diverse questions[4]. They help healthcare professionals interpret images and build trust in AI predictions. However, achieving high accuracy while maintaining logical consistency, especially in tasks like DME staging, remains a significant challenge[5],[6]. Inconsistent answers can undermine the reliability of the models in clinical practice[7].

The novelty of our approach lies in a consistency-enforcing mechanism that aligns answers to perception-based and reasoning-based questions, without relying on external datasets or introducing performance trade-offs. This mechanism ensures that the model provides accurate and consistent responses, thereby enhancing its interpretability and trustworthiness in critical medical imaging tasks. Existing methods to improve consistency often reduce accuracy[8],[9] limiting performance in real-world applications, where both are crucial[10].

The primary objective of this work is to achieve higher accuracy compared to ResNet101 while ensuring consistent and reliable performance in question-answering tasks. To address this, we propose a VQA model that balances high accuracy with logical consistency between related questions. As shown in Fig.1, the grade 0 indicates a healthy retina without exudates, grade 1 indicates exudates in the peripheral retina, grade 2 indicates severe DME with exudates in the macular region, and hard exudates are lipid deposits in the retina that indicate the severity of vascular leakage and are used to grade the disease. The proposed system helps to assess the severity of DME for accurate diagnosis and treatment.

The remainder of this paper is organized into five sections. Section 2 reviews related work, focusing on enhancing VQA systems through improved datasets, fine-grained feature extraction, and inter-modal reasoning, with an emphasis on

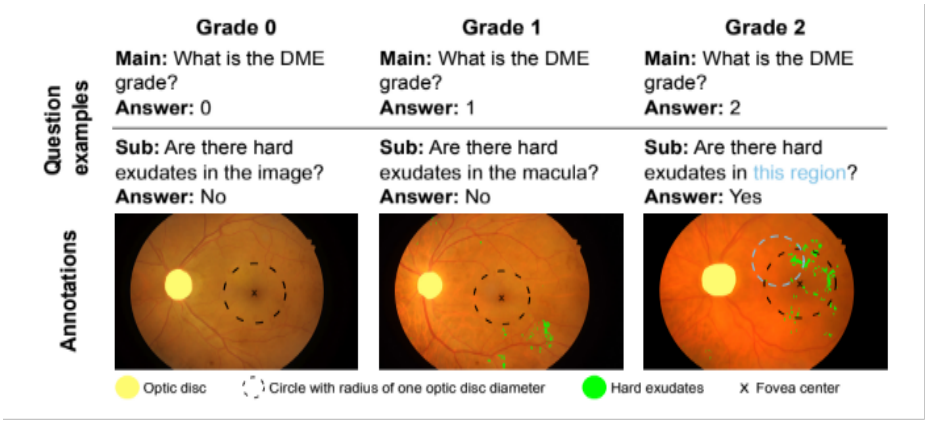


Fig. 1. Illustration of DME grading system with main and sub-questions: Main questions assess the overall DME grade, while sub-questions focus on specific regions, such as the macula or peripheral retina, with annotated features like optic disc and hard exudates for enhanced interpretability[11]

accuracy and consistency in medical diagnostics. Section 3 outlines the proposed methodology including the multimodal pipeline for pre-processing, feature extraction, and fusion techniques for better accuracy,consistency,and reliability. Section 4 presents the results and analysis highlighting the VGG16 model's superior performance in DME grading and its consistent predictions across similar questions, demonstrating its effectiveness in multimodal data handling. Section 5 is about the conclusion and outlines how the model uses VGG16 to analyze medical images and combinesWord2Vec with LSTM to accurately and consistently answer complex medical questions, outperforming ResNet101 in DME grading.Finally, Section 6 summarizes the key findings and suggests directions for future research.

2 Related Work

VQA is used in fields like assistive technologies and medical imaging, where it helps to answer questions regarding images by analyzing image data and understanding text. Older VQA methods use image processing techniques like extracting features with ResNet and text analysis tools such as GloVe and LSTMs to make sense of written questions. These methods struggle in handling intricate relationships between text-based queries and image data, especially in crucial fields like medical diagnostics, while still being effective for straightforward tasks. Limitations such as limited data availability, variations in answers, and challenges in detecting fine details further reduce their reliability and practicality in critical applications [12]. Improvements in focus-based techniques which includes co-attention and self-attention methods, have enhanced the VQA accuracy by focusing key areas of an image and relevant textual components [13].

However, existing methods still struggle to identify subtle features in images, such as small or overlapping objects, which is essential for medical imaging tasks such as the detection of diabetic macular edema (DME) in fundus images [14]. To address these issues, this research suggests enhancing dataset creation using automated and semi-automated methods to produce diverse, high-quality data while reducing noise. It also incorporates refined focus-based methods and innovative attention mechanisms to maintain consistency across related queries, while effective combination techniques boost the interaction between visual and textual elements [15]. In medical imaging, VQA systems can aid clinicians by offering clear and precise information to support the diagnosis of conditions such as DME. Beyond healthcare, these technologies can improve assistive tools for visually impaired individuals and facilitate advanced human-computer interactions. Future efforts will hierarchize the development of standardized data sets for medical imaging and enhance interpretability through transparent methods, setting new benchmarks for applications in diverse areas [16].

3 Methodology

The aim of the proposed methodology was to increase the accuracy of the VQA model by smoothly integrating visual and textual data. This approach was divided into several essential phases : The questions and images were preprocessed , meaningful features were extracted and these features were combined to provide reliable predictions. By using pre-trained models and advanced techniques, each phase was meticulously designed to ensure that the text and image data work together effectively.

3.1 Data Preprocessing

As first step, the model accepts two inputs ,an image and a textual query. To make sure both inputs were consistent with the model's later stages, they were processed independently. In order to maintain consistency throughout text preprocessing, the input questions were tokenized into individual words and changed to lowercase using NLTK. The semantic relationships between words were subsequently recorded by mapping each word to a 300-dimensional vector using the Word2Vec embedding model .The sequence was padded or trimmed to a fixed length of 40 tokens in order to standardize the input length. This guarantees that the input sequence (n = 40) has a consistent length. Zero-padding is used for sequences with fewer than 40 tokens:

$$\boldsymbol{Q} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n, \mathbf{0}, \ldots, \mathbf{0}] \quad \text{if } n < 40 \tag{1}$$

where the word vector for the token i-th in the sequence was represented by the vector $\boldsymbol{v}_i$.The Input sequences were represented by their length n. If a sequence contains fewer than 40 tokens, it was supplemented with a zero vector $\mathbf{0}$ to reach the required length of 40.

For sequences with more than 40 tokens, truncation was used. The sequence was cut down to a maximum of 40 tokens if it contains more than 40 tokens. With this, longer sequences can be reduced to a fixed size, enabling the model to handle them efficiently without incurring excessive computational costs.

$$\boldsymbol{Q} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{40}] \quad \text{if } n > 40 \tag{2}$$

where the i-th token in the sequence is represented by its word vector $\boldsymbol{v}_i$. When the sequence length goes beyond 40 tokens, it is cut down to the first 40 tokens.

Simultaneously, the visual data passed through preprocessing to ensure their compatibility with the VGG16 model[17]. By resizing the images to a larger resolution of 448x448 pixels, more intricate details from the original image can be retained. Due to this higher resolution, the model can capture finer features, which enhances its predictive accuracy. To be consistent with the VGG16 preprocessing pipeline, the images initially loaded in BGR format were converted to RGB format. Data standardization for efficient feature extraction and optimal model performance.Textual and visual data are now ready for effective feature extraction and fusion, facilitating seamless integration for precise predictions.

3.2 Feature extraction and Concatenation

To capture meaningful representations for each modality, feature extraction was carried out separately for both textual and visual inputs. For textual features,the preprocessed text data are fed into a LSTM network, which excels at capturing the sequential relationship between words in a question. The output of the LSTM was then refined through dense layers with ReLU activation functions, enabling the model to learn complex patterns. Dropout regularization was applied to prevent overfitting , ensuring that the textual embeddings capture both the contextual and semantic nuances effectively.

For visual features, the preprocessed images were fed into the VGG16 model, a pre-trained convolutional neural network excludes the fully connected layers (classification head) enabling the use of its convolutional layers exclusively for extracting high-level image features without performing classification. These feature maps, with dimensions of $14\times14\times512$, encode spatial and semantic information about the visual input.

To align these visual features with the textual embeddings, dimensionality reduction was performed using a combination of convolutional and dense layers, resulting in a 21×300 representation. The dimensionality reduction step simplifies the high-dimensional visual data, making it easier for the model to handle and more efficient to process. By reducing the visual data to a size that matches the textual data, it allows both types of information to work together smoothly. This makes it easier for the model to combine the image and text, helping it make more accurate predictions. Early fusion,late fusion and hybrid fusion are techniques for combining features or information from multiple modalities in a

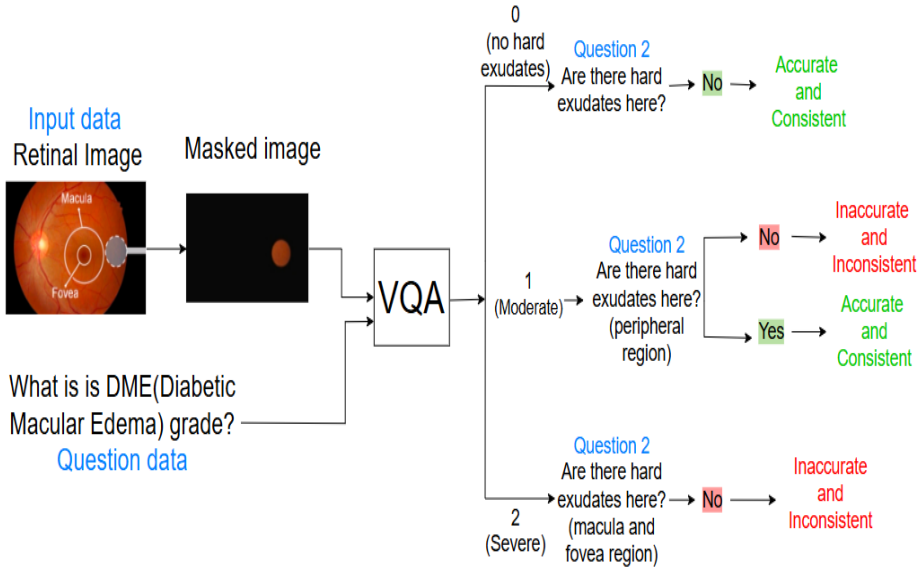


Fig. 2. Illustration of the VQA process for assessing the (DME)grade.The system analyzes a masked retinal image and answers a series of questions regarding the presence and severity of hard exudates in different regions, thereby providing outputs with varying levels of accuracy and consistency.

model. Late fusion through feature concatenation was employed to merge textual and visual features. This approach preserves the strengths of each modality while creating a unified representation for prediction.

3.3 Model Training

The training process included the optimization of visual and textual features combined to ensure precise predictions. After the visual and textual embeddings are concatenated to create a unified representation in a process of feature fusion, the resulting multimodal feature embeddings undergo further processing by being passed through dense layers. The layers assist the model in grasping intricate relationships and patterns linking visual and textual data.

To categorize the fused features into one of the 476 predefined answer categories, a softmax activation function was utilized at the output layer. The model was trained with the categorical cross-entropy loss function, which quantifies the difference between the predicted and actual distributions of answers. The Adam optimizer was employed to optimize the training process due to its capability of dynamically adjusting learning rates based on gradient updates, which ensures faster convergence. Moreover, there was a ReduceLROnPlateau callback that keeps track of validation performance throughout training and lowers the learning rate if there was no improvement over a specified number of epochs.

This assists in adjusting the model with precision and preventing stagnation throughout the optimization process. Training was carried out over 49 epochs, with hyper-parameter tuning that involved modifying the batch size and learning rate. Consequently, the model demonstrates excellent performance, attaining a training accuracy of 96% and a validation accuracy of 89%. The model is able to effectively predict answers based on multimodal inputs due to this rigorous training process, which ensures good generalization.

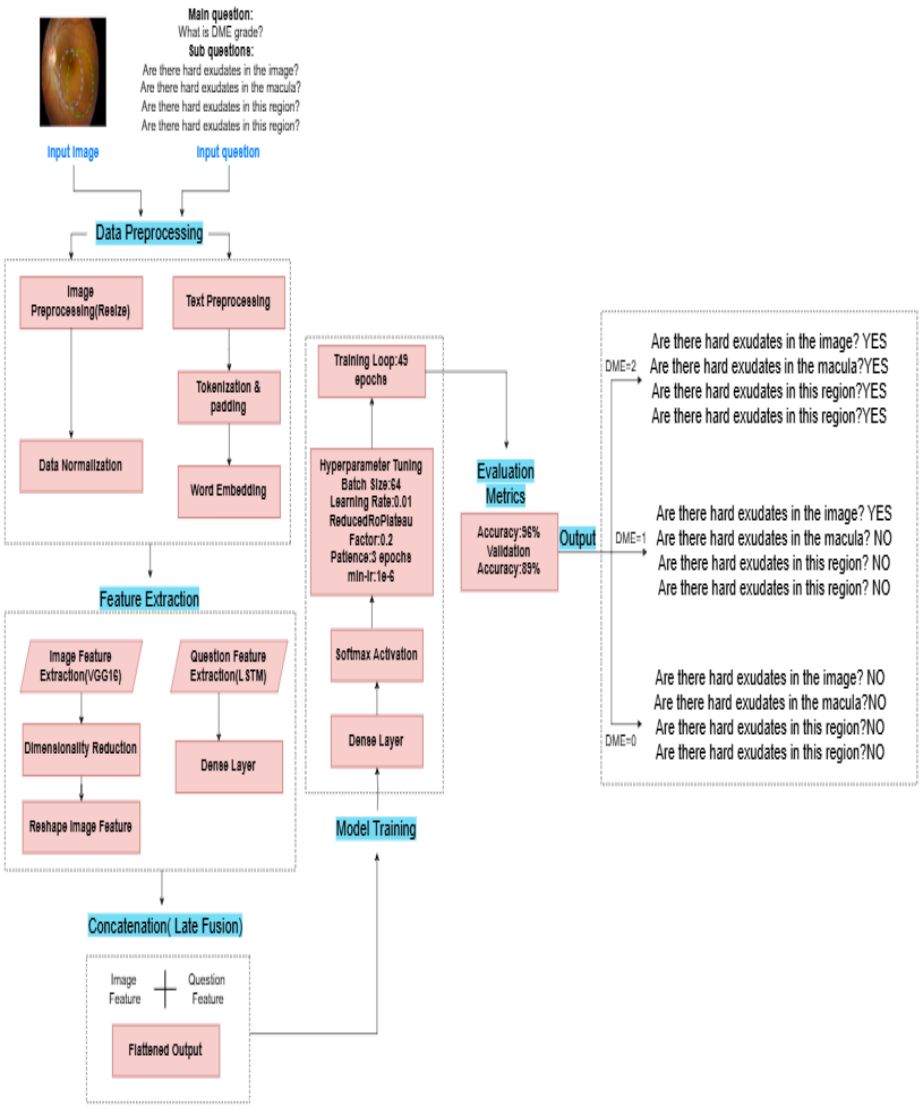


Fig. 3. Multimodal Pipeline Overview

4 Results and Discussions

4.1 Dataset Description

This dataset was developed to evaluate the reliability of VQA models with regards to the analysis of DME in fundus photographs. It incorporates images from the IDRiD and eOphta datasets, which are well-known in retinal imagery analyses. The dataset consists of diseased and not diseased retinal images, each with a specific question related to the grade of DME and the presence of hard exudates. The dataset consists of 433 images and 9779 question-answer pairs in the training set, 112 images and 2380 question-answer pairs in the validation set, and 134 images with 1311 question-answer pairs in the test set.[18].

The main focus was to examine if the VQA models took similar queries related to similar images and provided consistent responses. A distinguishing feature of this dataset are the region-based questions that allow the model to view specific regions of the image and query them.

Question Types: Automatically generated questions are derived from DME grade annotations with a focus on the segmentation of hard exudates. Included in the questions are the following: For the last two types of question, associated region masks were provided, where the specific size and location of the region is provided. Questions regarding the presence of the optic disc are not germane to the task of DME grade assignment.

Table 1. List of Questions For Diabetic Macular Edema (DME) assessment.

S.No.	Question Type
1	What is the diabetic macular edema grade for this image?
2	Are there hard exudates in this image?
3	Are there hard exudates in the fovea?
4	Are there hard exudates in this region?
5	Are there optic discs in this region?

4.2 Advantages of the Proposed VQA Model

The model has a number of weaker points compared to the proposed model but it is explained that the VQA model is actually more beneficial because it is easier to use in domain specific tasks. Using VGG16 as a visual feature extraction model, the Visual Domain Adaption Model was able to extract high-level image features with less effort and perform better in specialized domains, such as medical imaging. It does not perform as well as other models like ResNet101 because it needs more memory and processing resources to achieve strong results, but it greatly outperforms lower-quality models.

The integration of Word2Vec with LSTM enables the model to process text better by capturing semantic meaning, leading to better answers. Instead of ensuring text comprehension, this method enables deeper understanding of the text.

The model merges visual and textual modalities using a basic concatenation approach in multimodal fusion. Despite being easier than more sophisticated techniques, this method meaningfully fuses both modalities so that they contribute to the predictions. Additionally, the model responds to questions that are related, achieving natural coherence in answering without the need for specific loss functions, making it robust and context aware.

To put things in perspective, the model is a compelling alternative to conventional approaches in regards to highly specialized datasets because of its flexibility, efficiency, and adaptability.

4.3 Performance Metrics

The accuracy for both training and validation improved consistently over the five epochs for the VGG16 based model. The training accuracy reached 96.88% at epoch 49, demonstrating good learning. The validation accuracy also increased, reaching 89. 17% at epoch 48 and then settling at 87.52%. This tells us that the model is able to generalize to new samples while also learning from the provided data.

The VQA model proves effective and adaptable, particularly for the dataset at hand. This is done by employing VGG16, a convolutional neural network architecture that is meant for a specific dataset, instead of ResNet101 which is suitable for general tasks. Using VGG16, the model captures the salient features of the dataset with high precision.

Word2Vec and LSTM were integrated to analyze and understand text data. This method has helped the model to get the meaning of the inputs of semantic text with better precision. Unlike other more sophisticated approaches, this method makes it easier to integrate features of images and text while still being very effective. The simplicity of this method increases the interpretability and decreases the computational cost, making it possible to process information faster without losing accuracy.

One additional advantage of the new model is that it provides consistent outputs without having predefined rules or constraints for different tasks and questions. The design philosophy of the model and its implementation is based on a holistic approach that focuses on the system's performance with domain specific tasks, which is why is is able to outperform more generic methods such as ResNet101. Thus, the model was designed for high precision and responsiveness, and using simple methods makes it very efficient.

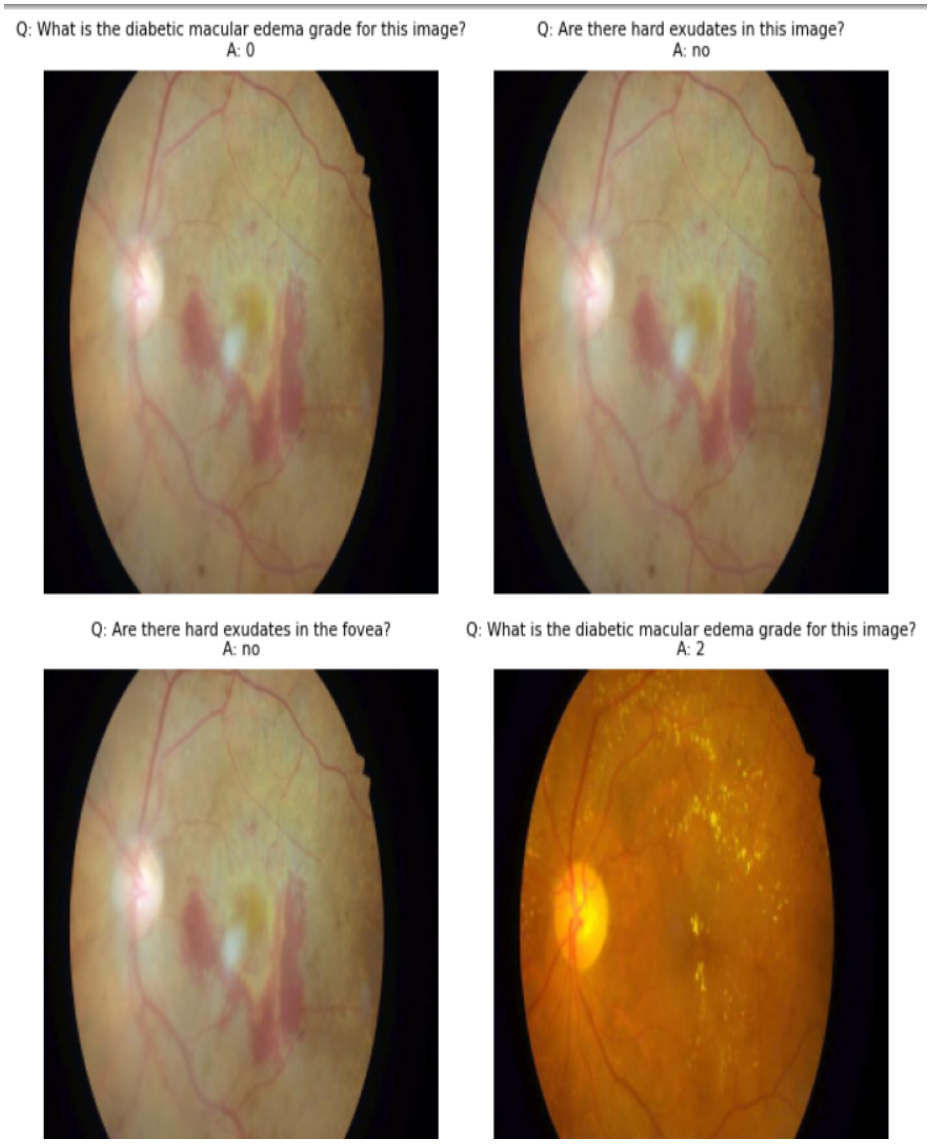
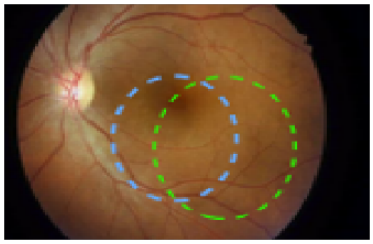


Fig. 4. VQA model is able to generate consistent answers across multiple scenarios, as illustrated in the provided examples. Despite not using formal consistency metrics, the responses remain stable and accurate when evaluating similar features, such as the diabetic macular edema grade and the presence of hard exudates. This demonstrates the model's reliability in maintaining uniform interpretation of similar visual and contextual patterns, ensuring dependable outcomes.

Table 2. Evaluation of responses for DME grading and detection of hard exudates, comparing ground truth values with model predictions.



Question	Type	Ground Value	Predicted Output
What is the DME grade?	MAIN	0	0
Are there hard exudates in the image?	SUB	YES	YES
Are there hard exudates in the macula?	SUB	NO	NO
Are there hard exudates in this region?	SUB	YES	YES
Are there hard exudates in this region?	SUB	YES	YES

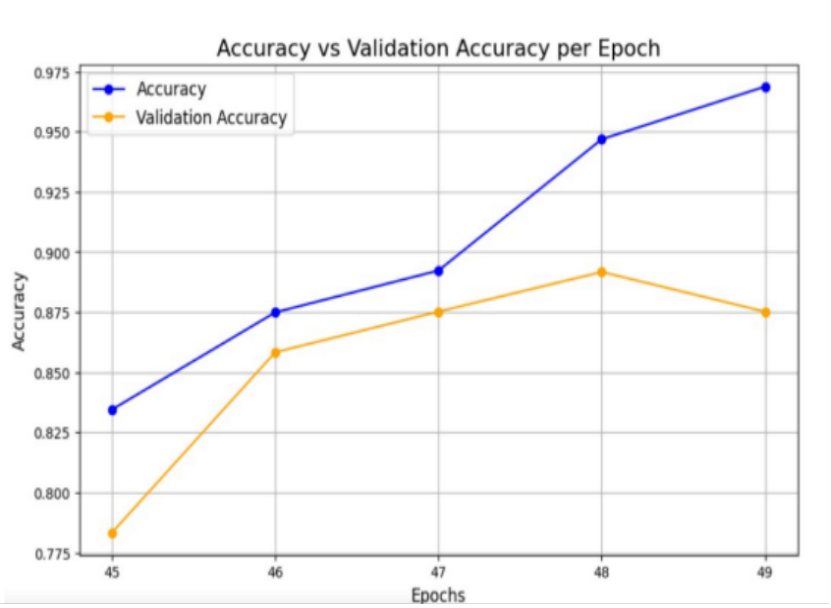


Fig. 5. The graph illustrates the progression of training accuracy and validation accuracy over epochs for the VGG16-based model. The training accuracy (blue line) consistently improves, peaking at 96.88% at epoch 49. The validation accuracy (orange line) initially increases, reaching a maximum of 89.17% at epoch 48, before slightly stabilizing at 87.52%.

5 Conclusion

The objective of the study was achieved with a training accuracy of 96.88% and a validation accuracy of 87.52%, and the performance was consistent across the board. The VGG16 VQA model outperformed the ResNet101 model in the specialization of medical imaging, especially in the grading of diabetic macular edema (DME), which confirms the possibility of VGG16 being used in these imaging applications because it appears to be very good at visual feature extraction. Furthermore, the implementation of Word2Vec for natural language processing together with LSTM for question answering in sequence comprehension greatly enhanced the model's comprehension and responsiveness, which was markedly consistent in complex medical questioning. These aspects contribute to the fact that effective visual information and language comprehension processing must be integrated to formulate dependable clinical solutions.

6 Future Work

Future work could explore the introduction of formal consistency metrics and the use of advanced technologies, such as transformer-based architectures (e.g., BERT or Vision Transformers), to improve model performance and reliability. Incorporating attention mechanisms, such as self-attention, along with knowledge graphs or external ontologies for better contextual understanding, may further improve the logical consistency of model output. In addition, expanding the scope of the model to address a wider range of medical imaging tasks and clinical questions would broaden its applications in healthcare. These improvements could pave the way for more consistent, interpretable, and practical solutions in medical decision making.

References

1. Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Pengjun Xie, Philip S. Yu, Fei Huang, and Jingren Zhou. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent, 2024.
2. Nghia Hieu Nguyen, Duong T.D. Vo, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. Openvivqa: Task, dataset, and multimodal fusion models for visual question answering in vietnamese. *Information Fusion*, 100:101868, December 2023.
3. Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at vqa models: Introspecting vqa models with sub-questions, 2020.
4. Haifan Gong, Guanqi Chen, Sishuo Liu, Yizhou Yu, and Guanbin Li. Cross-modal self-attention with multi-task pre-training for medical visual question answering, 2021.
5. Xixi Ga, Wenjie Liu, Tongyu Zhu, Shan Kou, Meishen Liu, and Yue Hu. Evaluating robustness and diversity in visual question answering using multimodal large language models. 2024.

6. Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering, 2020.
7. Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence*, 7, November 2024.
8. Vatsal Goel, Mohit Chandak, Ashish Anand, and Prithwijit Guha. Iq-vqa: Intelligent visual question answering. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pages 357–370. Springer, 2021.
9. Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation, 2019.
10. Yuanyuan Yuan, Shuai Wang, Mingyue Jiang, and Tsong Yueh Chen. Perception matters: Detecting perception failures of vqa models using metamorphic testing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16903–16912, 2021.
11. Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. Consistency-preserving visual question answering in medical imaging, 2022.
12. Amit Gangwal, Azim Ansari, Iqrar Ahmad, Abul Kalam Azad, and Wan Mohd Azizi Wan Sulaiman. Current strategies to address data scarcity in artificial intelligence-based drug discovery: A comprehensive review. *Computers in Biology and Medicine*, 179:108734, 2024.
13. Vibhashree B. S, Nisarga Kamble, Sagarika Karamadi, Sneha Varur, and Padmashree D Desai. Beyond words: Exploring co-attention with bert in visual question answering. In *2024 5th International Conference for Emerging Technology (INCET)*, pages 1–6, 2024.
14. Minh H. Vu, Tommy Lofstedt, Tufve Nyholm, and Raphael Sznitman. A question-centric model for visual question answering in medical imaging. *IEEE Transactions on Medical Imaging*, 39(9):2856–2868, September 2020.
15. Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE Transactions on Medical Imaging*, 42(5):1532–1545, 2023.
16. Rajat Subraya Gaonkar, V A Pruthvi, L Prem Kumar, Rohan Madan Ghodake, M J Raghavendra, and B Niranjana Krupa. Fine-grained feature extraction from indoor data to enhance visual question answering. In *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 895–902, 2023.
17. Satwik Kulkarni, Sitanshu Hallad, Tanvi Bhujannavar, Abhishek S. Masur, Shankru Guggari, Uday Kulkarni, and S. M. Meena. Elpdi: A novel ensemble learning approach for pulmonary disease identification. In Aditya Kumar Singh Pundir, Anupam Yadav, and Swagatam Das, editors, *Recent Trends in Communication and Intelligent Systems*, pages 49–60, Singapore, 2023. Springer Nature Singapore.
18. Sergio Tascon Morales, Pablo Márquez-Neila, and Raphael Sznitman. Diabetic macular edema vqa dataset, June 2022.