**Clinician's Diagnosis with AI-Powered Visual Question Answering: A Resource-Constrained Approach for Medical Imaging**

**Table of Contents**

1. **Introduction**  
   1.1. Problem Definition  
     • Overview of early disease detection in medical imaging  
     • Importance of timely diagnosis and its impact on patient outcomes  
   1.2. Significance of the Project  
     • Clinical relevance and real-world impact  
     • The value of integrating VQA for interactive diagnostics  
   1.3. Objectives and Scope  
     • Specific goals of the project  
     • Justification for selected AI techniques given hardware constraints  
   1.4. Report Structure  
     • Brief overview of the sections covered in the report
2. **Literature Review**  
   2.1. Overview of Existing Solutions  
     • Review of traditional and deep learning-based approaches in medical image analysis  
     • Discussion of current methods in early disease detection.  
   2.2. Visual Question Answering (VQA) in healthcare  
     • Existing VQA models and datasets.  
     • Integration challenges and benefits of VQA in a clinical context  
   2.3. Comparison of Methodologies  
     • Advantages and limitations of transfer learning versus training lightweight models on limited hardware  
     • Comparative analysis of segmentation approaches.  
     • Analysis of multimodal fusion techniques in VQA  
   2.4. Summary of Literature Findings  
     • Key insights that justify your project’s methodology  
     • Identified research gaps and how your work aims to address them
3. **Project Methodology**  
   3.1. Overall Approach  
     • High-level description of the integrated system for classification, segmentation, and VQA  
   3.2. Data Selection and Preprocessing  
     • Description of selected datasets.  
     • Preprocessing steps: normalization, resizing, augmentation, and segmentation preparation  
   3.3. Feature Extraction and Model Fine-Tuning  
     • Explanation of transfer learning and fine-tuning CNNs.  
     • Rationale for using lightweight models on CPU-limited hardware  
   3.4. Visual Question Answering Module  
     • Detailed breakdown of the VQA module development  
     • Synthetic query generation  
     • Fusion techniques for integrating visual and textual data  
   3.5. Integration Strategy  
     • How the classification, segmentation, and VQA components will be combined  
     • Data flow and system architecture overview
4. **Implementation Details**  
   4.1. Software and Hardware Environment  
     • Detailed list of libraries, frameworks, and tools used   
     • Discussion of hardware limitations and strategies for efficient CPU-only training  
   4.2. Code Architecture and Workflow  
     • Code structure overview with key modules and functions  
     • Flowcharts and diagrams illustrating the system pipeline  
   4.3. User Interface Design  
     • Description of the UI for uploading images and querying the system
5. **Evaluation and Analysis**  
   5.1. Evaluation Metrics  
     • Metrics for classification  
     • Metrics for segmentation  
     • VQA evaluation criteria  
   5.2. Experimental Setup  
     • Description of the testing dataset and experiment configuration  
     • Details of the validation process  
   5.3. Performance Analysis  
     • Comparison of results with baseline methods  
     • Discussion of improvements achieved by your integrated approach  
   5.4. Error Analysis and Limitations  
     • Identification of common errors and potential causes  
     • Discussion on the limitations imposed by hardware constraints and small datasets
6. **Ethical Considerations**  
   6.1. Data Privacy and Security  
     • Discussion on handling sensitive medical data  
   6.2. Bias and Fairness  
     • Analysis of potential biases in training data and model predictions  
   6.3. Responsible AI Deployment  
     • Consideration of explainability and transparency in clinical applications  
     • Discussion on safeguarding against over-reliance on automated diagnostics
7. **Conclusion and Future Work**  
   7.1. Summary of Key Findings  
     • Recap of the system’s achievements and contributions  
   7.2. Project Contributions  
     • Highlighting novelty, such as the integration of lightweight models and VQA on limited hardware  
   7.3. Potential Improvements  
     • Areas for further optimization   
   7.4. Future Extensions  
     • Recommendations for extending the system’s functionalities and deployment in real-world settings
8. **References**  
   8.1. Cited Papers  
     • Full bibliographic details of the papers reviewed and referenced

**INTRODUCTION**

* 1. **Problem Definition**

Early disease detection in medical imaging is a challenging and timely task in modern healthcare diagnosis as it plays an important role in improving patient outcomes, reducing treatment costs, and enabling a more proactive clinical response. In medical imaging, like chest X-rays, MRI scans, or dermatoscopic images, accurately detecting and describing minor anomalies, including early-stage tumors or abnormal lesions, is very important. Even slight inaccuracies in segmentation can result in misdiagnosis or delay in treatment, as noted by Zhou et al., in their work on UNet++, where they emphasize that marginal errors in segmentation may obscure important features such as the little spiculation of nodules, which can indicate malignancy.

The problem increased with challenges of data scarcity and high variability in medical images. Traditional image analysis techniques that rely on handcrafted features often fall short of capturing the complex patterns needed for reliable diagnosis. Therefore, deep learning methods like the convolutional neural network (CNNs) are known as a powerful tool for this domain. Moreover, as the demand for explainable and interpretable AI solutions increases in clinical practice, there is a pressing need to integrate mechanisms that provide good diagnostic accuracy and allow clinicians to understand the model's decisions. This is where Visual Question Answering (VQA) comes into play by enabling an interactive diagnostic process where clinicians can ask targeted questions about an image and receive context-sensitive, interpretable responses. The system can improve the transparency and usability of AI-based diagnostic tools.

The main problem addressed in this project is the development of an interactive AI diagnostic system for early disease detection that combines robust image segmentation and classification with a VQA module. This integration aims to deal with the challenges of achieving high diagnostic accuracy under resource constraints and encouraging clinical trust through improved explainability.

**1.2 Significance of the Project**

This project directly addresses the challenges faced in clinical practice. In many healthcare settings, early and accurate diagnosis through medical imaging can make a big difference in how a patient is treated. Quick and understandable diagnosis saves lives and reduces the costs and complications associated with delayed treatments due to late identification of the disease in diagnosis. For example, catching a lung nodule at an early stage might help prevent the disease from becoming more serious, which could lead to better outcomes for patients.

In addition to improving diagnostic accuracy, the project aims to make the process more interactive using Visual Question Answering (VQA) by allowing clinicians to ask specific questions about a medical image, like Where exactly is the lesion? or how big is the abnormal area? The project is designed to improve diagnostic accuracy and empower healthcare professionals with clearer, more engaging insights from medical images, which can lead to better patient care, enhanced workflow, and ultimately a more reliable and transparent diagnostic process in real-world healthcare environments.

**1.3 Objectives and Scope**

The main goal of this project is to design, develop, and evaluate an AI-based system that can detect early signs of diseases from medical images and offer interactive diagnostic support through visual question answering (VQA). The project aims to:

1. Develop a deep learning model for medical image classification and segmentation that accurately identifies abnormal regions in diagnostic images.
2. Integrate a VQA module that enables clinicians to ask targeted questions about specific image regions and receive meaningful, interpretable answers.
3. Ensure the model remains lightweight and efficient so that it can be trained and run on a typical CPU-based laptop to address hardware constraints without sacrificing performance.

**1.4 Report Structure**

This report is structured into sections to document the project’s development and findings:

1. **Introduction:**  
   This section provides the problem, its significance, and sets out the project’s objectives and scope.
2. **Literature Review:**  
   This section examines existing solutions, models, and technologies related to medical image analysis and VQA.
3. **Project Methodology:**  
   This section shows a detailed step-by-step approach used in this project, including dataset selection, data preprocessing, model selection, and the techniques used for building the models and developing the VQA module.
4. **Implementation:**  
   This section describes the system’s architecture and the integration of its components with code snippets, flowcharts, and diagrams.
5. **Evaluation and Analysis:**  
   This section presents the metrics and methods used to assess the performance of the developed system.
6. **Ethical Considerations:**  
   This section addresses issues related to data privacy, biases in model training, and the broader ethical implications of deploying such an AI-based diagnostic system in clinical settings.
7. **Conclusion and Future Work:**  
   This section summarizes the key findings and contributions of the project and suggests directions for future improvements and further research, particularly on scaling the solution or enhancing its interactive capabilities.

**LITERATURE REVIEW**

**2.1. Overview of Existing Solutions**

The field of medical image analysis has notably changed over the past few years, transitioning from traditional methods to advanced deep learning techniques. Historically, medical image analysis relied on manual feature extraction techniques combined with classical machine learning algorithms. Methods such as histogram analysis, edge detection, and texture analysis were commonly employed to identify disease markers in images (Kumar et al., 2012). For instance, the use of histograms of oriented gradients (HOG) features for tumor detection in mammograms exemplifies traditional approaches (Dalal & Triggs, 2005). However, these methods often faced challenges related to variability in image quality and the subjective nature of manual feature selection, leading to inconsistent results across different studies.

The introduction of convolutional neural networks (CNNs) marked a fundamental change in medical image analysis. CNNs automate the feature extraction process, allowing for more robust and accurate analysis of complex medical images. Ronneberger et al. (2015) introduced U-Net, a deep learning architecture specifically designed for biomedical image segmentation. U-Net's encoder-decoder structure, with skip connections, enables the model to capture both low-level and high-level features, making it particularly effective for tasks such as cell segmentation in histopathology images. Recent advancements have further refined these architectures. For example, UNet++ by Zhou et al. (2018) enhances the original U-Net by incorporating nested skip pathways and deep supervision, which improves segmentation accuracy and reduces the semantic gap between encoder and decoder features. This is particularly relevant for early disease detection, as accurate segmentation of abnormal regions can significantly impact diagnostic outcomes.

In the context of early disease detection, CNNs have been widely adopted for various applications, including lung cancer detection from CT scans and diabetic retinopathy classification from retinal images. For instance, Rajpurkar et al. (2017) demonstrated that a deep learning model could achieve high accuracy in pneumonia detection from chest X-rays, outperforming traditional methods. Similarly, DenseNet has shown promise in classifying skin lesions, leveraging its feature reuse capabilities to improve performance on small datasets (Huang et al., 2017).The integration of transfer learning has also become a common practice in medical image analysis. By utilizing pre-trained models on large datasets like ImageNet, researchers can fine-tune these models for specific medical tasks, effectively addressing the data scarcity problem often encountered in healthcare (Kim et al., 2021). This approach not only saves time and computational resources but also enhances model performance, making it a valuable strategy for early disease detection.

**2.2. Visual Question Answering (VQA) in Healthcare**

Visual Question Answering (VQA) systems enable machines to answer natural language questions about visual content and have been successful in healthcare as tools to enhance diagnostic workflows. Early VQA frameworks in healthcare focused on combining image features extracted by convolutional neural networks (CNNs) with text embeddings from recurrent neural networks (RNNs). VQA-RAD (Lau et al., 2018) is one of the first medical VQA datasets, containing 3,515 radiology images paired with question-answer sets like Is there a tumor in the brain MRI? Models trained on VQA-RAD used CNN-LSTM architectures, achieving 67% accuracy, but struggled with complex spatial reasoning (Lau et al., 2018).

Recent advancements have been made in transformer-based architectures to improve multimodal fusion. For instance, PMC-VQA (Liu et al., 2021) has a dataset with 227,000 medical image-question pairs, enabling models like MMBERT (Khare et al., 2021), a multimodal BERT that demonstrated high performance in answering pathology-related questions by jointly encoding visual and textual features.

**2.2.1. Integration Challenges in Clinical Contexts**

Aside from the progress made in deploying VQA systems in healthcare, there are still some challenges:

1. Interpretability: Clinicians require explanations for answers, especially in high-stakes scenarios. While models like ViLT (Kim et al., 2021) generate attention maps to highlight relevant image regions, these visualizations often lack clinical granularity (Adhikari et al., 2023).
2. Data Scarcity: Medical VQA datasets remain limited in size and diversity, for example the VQA-RAD covers only radiology, leaving gaps in dermatology or ophthalmology (Lau et al., 2018).
3. Computational Overhead: Transformer-based models, though accurate, demand GPU resources, conflicting with the project’s goal of CPU-compatible deployment (Dosovitskiy et al., 2021).

**2.2.1. Benefits of VQA in Clinical Workflows**

When VQA systems are effectively integrated, they can:

* Reduce Diagnostic Time: A study by Adhikari et al. (2023) showed that clinicians using a VQA prototype for lung X-ray analysis reduced interpretation time by 30% compared to traditional methods.
* Enhance Accessibility: VQA tools can guide non-specialists by answering targeted questions in resource-limited settings e.g., Is this mole malignant? (Khare et al., 2021).
* Support Training: VQA systems act as educational aids for medical students by providing instant feedback on image-based queries (Liu et al., 2021).

**2.3. Comparison of Methodologies**

This section evaluates the strengths and limitations of methodologies relevant to the project, including transfer learning versus lightweight model training, segmentation architectures, and multimodal fusion techniques in VQA.

**2.3.1. Transfer Learning vs. Lightweight Models**

Transfer learning involves fine-tuning pre-trained models on medical datasets and has become the basis of medical image analysis. Tajbakhsh et al. (2016) demonstrated that fine-tuning CNNs pre-trained on natural images like ImageNet improved skin lesion classification accuracy by 15% compared to training from scratch. However, these models mostly require more computational resources, making them unsuitable for deployment on CPU-only systems.

Lightweight architectures like MobileNet (Howard et al., 2017) address this by using depth-wise separable convolutions to reduce parameters. The study by Sandler et al. (2018) showed that MobileNet achieved 89% accuracy in diabetic retinopathy detection while using 75% fewer parameters than ResNet-101. However, lightweight models may sacrifice performance on complex tasks like segmentation, where spatial detail is important (Zhou et al., 2018).

**2.3.2. Segmentation Approaches: U-Net vs. U-Net++**

U-Net remains a benchmark for medical image segmentation due to its ability to localize abnormalities with minimal training data. Ronneberger et al. (2015) reported a 92% Dice score for neuronal structure segmentation using U-Net. However, its dependency on fixed skip connections can lead to rough feature maps in complex cases, such as overlapping lung nodules (Zhou et al., 2018).

UNet++ (Zhou et al., 2018) addresses this by introducing nested skip pathways and deep supervision. In experiments on liver tumor segmentation, UNet++ improved Dice scores by 8% over U-Net by reducing the semantic gap between encoder and decoder layers. Additionally, its architecture allows pruning during inference, making it adaptable for lightweight deployment, which is an advantage for the project’s hardware constraints.

**2.3.3. Multimodal Fusion in VQA**

Effective VQA systems require a strong fusion of visual and textual features. Early approaches like CNN-LSTM (Lau et al., 2018) simply link image and text embeddings together, resulting in limited spatial reasoning. For instance, models in the VQA-RAD dataset achieved only 61% accuracy for questions requiring localization, like where is the tumor?

Transformer-based fusion techniques, like ViLBERT (Lu et al., 2019) use cross-modal attention to align image regions with text tokens. This approach in the PMC-VQA dataset improved accuracy to 74% for complex queries (Liu et al., 2021). However, transformers are computationally intensive: ViLBERT requires 16GB of GPU memory, which conflicts with the project’s CPU-compatibility goal. Tang et al.'s (2023) recent work proposes dynamic token pruning to reduce computation by 40% without significant accuracy loss, offering a viable path for efficient VQA integration.

**2.4 Summary of Literature Findings**

U-Net (Ronneberger et al., 2015) created the basic work for accurate medical image segmentation, but its limitations in handling complex anatomical structures were addressed by UNet++ (Zhou et al., 2018). The UNet++ nested skip pathways and deep supervision improve segmentation accuracy by 8% in liver tumor tasks while enabling model pruning for lightweight deployment. Fine-tuning pre-trained CNNs like DenseNet (Huang et al., 2017) or ResNet significantly improves classification accuracy in medical tasks (Tajbakhsh et al., 2016). However, models like MobileNet (Howard et al., 2017) offer a viable compromise, reducing computational demands by 75% while maintaining competitive performance (Sandler et al., 2018).

While transformer-based models like ViLT (Kim et al., 2021) enhanced multimodal fusion in VQA, it’s computational cost limits clinical applications (Dosovitskiy et al., 2021). Recent work on dynamic token pruning (Tang et al., 2023) reduces inference overhead by 40%, offering a pathway to efficient VQA integration. Additionally, datasets like VQA-RAD (Lau et al., 2018) and PMC-VQA (Liu et al., 2021) provide benchmarks for training clinically relevant models.

In the domain of Visual Question Answering for healthcare, datasets like VQA-RAD have enabled the development of systems that combine image analysis with natural language processing to answer clinically relevant questions about medical images. However, existing approaches face a few limitations:

1. Most current medical VQA systems require high computational resources, limiting their accessibility in resource-constrained environments.
2. The integration of classification, segmentation, and VQA capabilities into a unified system remains underexplored.
3. Balancing model complexity with performance under hardware constraints presents significant challenges for real-world deployment.

This project addresses these gaps by developing an integrated, resource-efficient system that combines these components while trying to maintain clinical utility and diagnostic accuracy.

**PROJECT METHODOLOGY**

**3.1. Overall Approach**

This project aims to create an integrated AI system designed for early disease detection in medical images while also providing an interactive diagnostic experience through Visual Question Answering (VQA). Given the resource limitations, this approach focuses on using lightweight yet effective deep learning models that can be fine-tuned on a CPU-based system. The proposed system integrates three core components, which are classification, segmentation, and visual question answering, into a unified framework for comprehensive medical image analysis. This integrated approach helps to perform mixed analysis on medical images, combining the strengths of each component to provide diagnostic support than any individual component could offer alone.

PUT A DIAGRAM FOR THE SYSTEM WORKFLOW

The system workflow begins with preprocessing input medical images to standardize quality and dimensions. These processed images are then analyzed through two parallel pathways:

1. The classification branch helps identify the presence and type of abnormalities using lightweight CNN architectures optimized for CPU deployment.
2. The segmentation branch helps precisely outline regions of interest within the image, highlighting potential disease markers.

The outputs from these branches, along with the original preprocessed image, serve as inputs to the VQA module, which processes natural language questions from clinicians and generates contextually relevant responses. This design enables clinicians to receive both automated analysis and interactive feedback through natural language queries about specific aspects of the medical image. The entire pipeline is optimized to operate efficiently on standard CPU hardware, with careful attention to model size, computational complexity, and memory usage. This approach ensures accessibility in various clinical settings without requiring specialized hardware resources.

**3.2. Data Selection and Preprocessing**

**3.2.1. Datasets:**

For this project, I will use the following publicly available datasets that cover a range of medical imaging modalities and pathologies:

1. NIH Chest X-ray Dataset: This dataset comprises over 112,120 frontal chest X-rays with 14 disease labels (Rajpurkar et al., 2017). It helps with the foundation for training and evaluating the classification component for thoracic pathologies.
2. ISIC Skin Lesion Dataset: I will leverage the International Skin Imaging Collaboration dataset, which contains 23,906 dermatoscopic images of skin lesions (benign/malignant) (Tajbakhsh et al., 2016).
3. VQA-RAD: This dataset contains 3,515 radiology images paired with 3,064 question-answer pairs created by clinicians, providing training data for the VQA component.
4. PMC-VQA: This dataset contains over 227,000 medical image-QA pairs for training the VQA module (Liu et al., 2021), providing training data for the VQA component.

**3.2.2. Preprocessing Steps:**

The preprocessing pipeline implements several strategies to enhance image quality and standardize inputs:

1. Normalization: All the images will be normalized to ensure pixel values fall within a consistent Min-max scaling (0–1) to address variability between images, to speed up convergence.
2. Resizing: All the images will be resized to uniform dimensions of 224×224 pixels for classification, 256×256 pixels for segmentation, for compatibility with the neural network architectures.
3. Noise Reduction: I will apply Gaussian filtering (σ=1.5) to the X-rays dataset to suppress unwanted signals, noise or distortions.
4. Data Augmentation: I will apply controlled augmentations to expand the dataset size and improve model generalization. Techniques like random rotations, horizontal flips, minor scaling, controlled brightness, and contrast adjustments.
5. Segmentation Mask Generation: For training the segmentation models, I will create binary masks to highlight regions of interest. For datasets without clear segmentation annotations, I will implement semi-supervised approaches to generate approximate masks using classification-based attention maps as initial guidance.

**3.3 Feature Extraction and Model Fine-Tuning**

Transfer learning will be applied to take advantage of pre-trained architectures, fine-tuning them for the specific medical imaging tasks.

**3.3.1. Classification Models**

For the classification component, several lightweight architectures will be optimized for CPU deployment:

1. MobileNetV2: This architecture employs inverted residuals and linear bottlenecks, achieving high efficiency through depth-wise separable convolutions. With only 3.5 million parameters, it offers an excellent balance between performance and computational cost. I will initialize the model with weights pre-trained on ImageNet and replace the final classification layer to match the target disease categories.
2. DenseNet-121: As an alternative, this architecture's dense connected pattern enables efficient feature reuse through direct connections between each layer and all following layers.

**3.3.2. Segmentation Models**

For the segmentation component, a modified U-Net architecture with key features of the implementation includes:

1. Lightweight Encoder: Instead of using the standard VGG-style encoder, a substitute MobileNetV2-based encoder is used to reduce computational requirements while maintaining feature extraction quality.
2. Skip Connections: Maintaining the skip connections between the encoder and decoder paths is important for preserving spatial information and enabling specific boundary localization.
3. Deep Supervision: Inspired by UNet++, deep supervision is achieved by adding extra loss functions at intermediate decoder levels to improve the gradient flow during training.

**Fine-Tuning Process:**

1. Freeze Early Layers: Start by freezing the initial layers of the network to capture basic features like edges and textures that are generalized. This reduces training time and computational cost.
2. Replace and Train Final Layers: Replace the original classification head with a new fully connected (FC) layer specific to the task and train this head on your dataset.
3. Incremental Unfreezing: Then, gradually unfreeze additional layers if the validation performance plateaus. This allows the network to adjust higher-level features without full-scale retraining.
4. Regularization: To prevent overfitting on the limited medical datasets, dropout (rate=0.2), weight decay, and early stopping based on validation performance will be applied.

To optimize the segmentation model for CPU deployment, efficiency-focused modifications like channel reduction in the decoder path to minimize memory requirements, efficient upsampling using bilinear interpolation instead of transposed convolutions, and strategic application of depth-wise separable convolutions at computational bottlenecks will be applied. These optimizations enable effective segmentation performance while maintaining compatibility with the hardware constraints.

**3.4 Visual Question Answering (VQA) Module**

Developing this module involves addressing interconnected challenges, like multimodal feature processing, alignment of visual and textual information, and generation of relevant clinical responses.

**3.4.1. Architecture Design**

The VQA module uses a multimodal architecture that processes image and text inputs in parallel before combining them:

1. Image Processing Branch: uses features from the intermediate layers of the classification and segmentation models and applies spatial attention to focus on regions relevant to the query, then generates multi-scale visual representations to capture both fine details and global context
2. Text Processing Branch: processes natural language questions using a simplified transformer encoder, which generates word embeddings that capture the semantic content of clinical queries and then implements medical vocabulary augmentation to handle domain-specific terminology
3. Multimodal Fusion: uses a hierarchical attention mechanism inspired by ViLT but optimized for CPU deployment to implement cross-modal attention that aligns to textual queries with relevant image regions and uses gated fusion to control information flow between modalities

**3.4.2. Synthetic Query Generation**

1. Template-Based Query Generation: Creating clinically relevant question templates that automatically fill templates with disease terms from the classification categories, generating questions that combine multiple attributes or regions.
2. Cross-modal Data Augmentation: Paraphrasing existing questions while preserving semantic intent and then generating variations of answers with equivalent clinical meaning.

**3.4.3. Fusion Techniques**

Effective integration of visual and textual information is important for generating accurate and relevant answers. A simplified version of multimodal attention that remains computationally efficient:

1. Question-Guided Visual Attention: The textual query directs attention to relevant regions in the visual features, creating question-specific visual representations.
2. Visual-Enhanced Text Representation: Visual features relevant to the question modify the text representation, enriching it with image-specific information.
3. Hierarchical Decision Making: Depending on the question type, the system routes the fused representation to specialized output heads optimized for each response type.

This fusion approach enables the system to ground its answers in specific image regions while maintaining the context of the clinical query, resulting in more specific and relevant responses.

**3.5 Integration Strategy**

The successful integration of classification, segmentation, and VQA components is important for creating a unified system that provides comprehensive diagnostic support. The integration strategy focuses on maintaining component modularity while enabling effective information flow between them.

**3.5.1 System Architecture**

The integrated system follows a modular pipeline architecture with shared feature extraction that enables efficient resource utilization while maintaining the specialized capabilities of each component:

1. Input Processing Layer: This layer accepts medical images in standard formats, then performs preprocessing and normalization. Also, it provides an interface for natural language queries.
2. Feature Extraction Backbone: This layer implements a shared encoder based on MobileNetV2 and generates multi-scale feature maps that feed into specialized heads for feature caching to avoid redundant computation.
   * Specialized Analysis Branches: This layer consists of the classification branch, that predicts disease categories, and the segmentation branch generates masks highlighting regions of interest
3. VQA Module: This layer receives features from both branches along with the processed query, and it implements multimodal fusion as described previously, after which it generates appropriate responses based on question type and feature analysis
4. Visualization and Interpretation Layer: This layer renders segmentation, and attention maps that format responses in relevant clinical terminology and also provides confidence indicators and supporting evidence.

**3.5.2 Data Flow**

The data flow through the integrated system follows a specific sequence optimized for both performance and interactive use. This sequential and interactive workflow enables clinicians to efficiently explore medical images, starting with automated analysis and progressing to targeted investigation of specific aspects or regions of interest:

1. Initial Analysis: Upon image upload, the system automatically performs classification and segmentation. The results are cached to enable rapid response to queries, and initial findings are presented to the user, highlighting potential areas of concern.
2. Interactive Querying: Users can submit natural language questions about the image, and the system processes these queries, using the cached analysis results, and responses are generated based on the combined information from all components.
3. Explanation Generation: For each response, the system provides supporting visual evidence and attention maps that highlight regions relevant to the question, with confidence scores that indicate the reliability of the provided answer.

**3.5.3 Component Communication**

1. Feature Sharing: classification and segmentation models share early-layer features to reduce redundant computation, both visual modules contribute features to the VQA component.

2. Attention Mapping: text queries guide attention in the visual processing pipeline, as segmentation masks help focus the VQA module on relevant image regions.

3. Caching Strategy: results from classification and segmentation are cached to enable rapid response to multiple queries about the same image. This reduces redundant computation and improves the interactive experience.

**IMPLEMENTATION DETAILS**

**4.1. Software and Hardware Environment**

The implementation of our AI-based diagnostic system utilizes a carefully selected technology stack optimized for performance on CPU-limited hardware:

**References**

* Kumar, A., & Gupta, A. (2012). A survey of medical image analysis techniques. International Journal of Computer Applications, 47(18), 1-6.
* Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 1, pp. 886-893). IEEE.
* Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI) (pp. 234-241). Springer.
* Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Laird, M., ... & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.
* Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2261-2269).
* Lau, J. J., Gayen, S., Abacha, A. B., & Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. Scientific Data, 5(1), 1-10.
* Liu, B., Zhan, L., Xu, L., Ma, L., Yang, Y., & Wu, X. (2021). PMC-VQA: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2105.13239.
* Adhikari, S., Poudel, M., & Li, X. (2023). Clinical VQA: Challenges and opportunities for deployable visual question answering in healthcare. Journal of Biomedical Informatics, 142, 104367.
* Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. arXiv preprint arXiv:1807.10165.
* Khare Yash, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, CV Jawahar. (2021). MMBERT: Multimodal BERT Pretraining for Improved Medical VQA. <https://arxiv.org/pdf/2104.01394>
* Kim Wonjae, Bokyung Son, Ildoo Kim (2021). ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. arXiv:2102.03334v2
* Dosovitskiy Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby (2021). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929v2.
* Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., & Gotway, M. B. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Transactions on Medical Imaging, 35(5), 1299-1312.
* Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
* Tang [Quan,](https://arxiv.org/search/cs?searchtype=author&query=Tang,+Q) [Bowen Zhang](https://arxiv.org/search/cs?searchtype=author&query=Zhang,+B), [Jiajun Liu](https://arxiv.org/search/cs?searchtype=author&query=Liu,+J), [Fagui Liu](https://arxiv.org/search/cs?searchtype=author&query=Liu,+F), [Yifan Liu](https://arxiv.org/search/cs?searchtype=author&query=Liu,+Y) (2023). Dynamic Token Pruning in Plain Vision Transformers for Semantic Segmentation. arXiv:2308.01045v2 [cs.CV].
* Sandler Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv:1801.04381v4