A Project Phase-I on

MediXpert: A Vision-Language System for Clinical Decision Support

Submitted to the Department of Computer Science & Engineering, GNITS in the partial fulfillment of the academic requirement for the award of B.Tech (CSE) under JNTUH

By

K.Yasaswitha (22251A0516)
A.Likhitha (22251A0520)
M.Srividya (22251A0522)
V.Rishika Reddy (22251A0564)

under the guidance of

Mrs. D.R. Nanda Devi
Assistant Professor

Department of Computer Science & Engineering
G. Narayanamma Institute of Technology & Science
(Autonomous) (For Women)
Shaikpet, Hyderabad - 500 104.

Affiliated to
Jawaharlal Nehru Technological University Hyderabad
Hyderabad – 500 085
October, 2025

G. Narayanamma Institute of Technology & Science
(Autonomous) (For Women)
Shaikpet, Hyderabad – 500 104.
Department of Computer Science & Engineering

Certificate

This is to certify that the Major Project report on "**MediXpert: A Vision-Language System for Clinical Decision Support**" is a bonafide work carried out by **K.Yasaswitha(22251A0516), A.Likhitha (22251A0520), M.Srividya (22251A0522), V.Rishika Reddy (22251A0564)** in the partial fulfillment for the award of B.Tech degree in Computer Science & Engineering, G. Narayanamma Institute of Technology & Science, Shaikpet, Hyderabad, affiliated to Jawaharlal Nehru Technological University, Hyderabad under our guidance and supervision.

The results embodied in the project work have not been submitted to any other University or Institute for the award of any degree or diploma.

Internal Guide
Mrs. D.R. Nanda Devi
Assistant Professor, CSE

Head of the Department
Dr. A.Sharada
Professor & Head
Department of CSE

External Examiner

Acknowledgements

We would like to express our sincere thanks to **Dr. K. Ramesh Reddy, Principal,** GNITS, for providing the working facilities in the college.

Our sincere thanks and gratitude to **Dr. M. Seetha,** Professor CSE , Dean R&D**, Dr. N. Kalyani,** Professor CSE , Dean I & I, GNITS, for all the timely support and valuable suggestions during the period of our project.

We extend our heartfelt gratitude to **Dr. A. Sharada,** Professor & Head, Department of Computer Science and Engineering, GNITS, for their unwavering support and invaluable guidance throughout our project, providing timely assistance and insightful suggestions.

We are extremely thankful to the overall coordinator of Project Phase-1 **Dr. D.V.Lalitha Parameswari,** Assoc. Professor, Dept. of CSE, GNITS for her encouragement and support throughout the project.

We are also extremely thankful to our project coordinators, **Dr. T. Rajesh,** Asst. Professor and **Dr. N. Anil Kumar,** Asst. Professor Department of CSE**,** GNITS for their encouragement and support throughout the project.

We are extremely thankful and indebted to our internal guide, **Mrs. D.R. Nanda Devi**, Assistant Professor, Department of CSE**,** GNITS for her constant guidance, encouragement and moral support throughout the project.

Finally, we would also like to thank all the faculty and staff of CSE Department who helped us directly or indirectly, parents and friends for their cooperation in completing the project work.

K.Yasaswitha (22251A0516)
A.Likhitha (22251A0520)
M.Srividya (22251A0522)
V.Rishika Reddy (22251A0564)

ABSTRACT

In today's healthcare systems, effective diagnosis often depends on a combination of patient-reported symptoms and the interpretation of diagnostic images such as chest X-rays, CT scans, or MRIs. However, most existing artificial intelligence (AI) tools focus on either one of these domains—language-based systems for processing patient queries (e.g., using transformers like BERT or GPT), or vision-based models for analyzing medical images (e.g., convolutional neural networks such as ResNet or EfficientNet)—without integratśing the two. This separation limits the diagnostic capabilities of AI solutions and fails to reflect the holistic approach that human clinicians follow when evaluating both verbal descriptions and visual data. The lack of a unified, intelligent system that can jointly process natural language and medical images results in fragmented care, delays in decision- making, and limited accessibility—especially in resource-constrained or remote environments.

To address this challenge, MediXpert is proposed as a unified vision-language AI assistant that bridges the gap between symptom-based dialogue and medical image understanding. The system allows users to describe symptoms in natural language while uploading medical images for analysis, interpreting both inputs to provide diagnostic guidance in a conversational format. It leverages advanced models such as transformers for language processing and convolutional neural networks for image analysis. By integrating natural language processing and computer vision, MediXpert aims to support clinical triage, early diagnosis, and patient education. This multimodal approach enhances decision-making by providing context-aware insights that single-modal systems often miss. It will be especially valuable for telemedicine platforms, virtual clinics, and remote diagnostics, augmenting healthcare professionals and expanding access to quality care.

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

Modern clinical decisions depend on both patient symptoms and diagnostic images, but most AI models handle them separately. This gap leads to slower and less accurate diagnoses. Multimodal AI systems aim to combine text and image data, allowing machines to interpret symptoms and visuals together, similar to how doctors think. These systems can provide more reliable and context-aware results. They help in early disease detection, reduce diagnostic errors, and improve efficiency. In low-resource settings, such AI tools can assist doctors by offering automated support and faster decision-making.

1.1 Background of the Study

In the modern healthcare environment, clinical decision-making increasingly depends on two primary types of information: patient symptoms and clinical imaging (e.g., X-rays, CT, MRI). Traditionally, artificial intelligence (AI) tools in medicine have addressed these two modalities in isolation: natural language processing (NLP) models focus on text (clinical notes, symptoms, history), while computer vision models focus on imaging. This separation, however, limits the ability of AI systems to mimic how clinicians integrate all available information into a coherent diagnosis.

As the complexity of diagnoses grows and resources remain constrained (particularly in low-resource settings), the need for integrated, multimodal decision support systems becomes more urgent. Vision-language models (VLMs) in the medical domain—also called medical vision-and-language models (MVLMs)—seek to bridge this gap by enabling simultaneous interpretation of visual and textual data. Recent research demonstrates that aligning image and text representations can improve tasks such as automatic report generation, medical visual question answering (Med-VQA), image-text retrieval, and multimodal diagnosis.

Modern clinical decisions rely on both patient symptoms and medical images, yet most AI models handle these separately. This separation causes inefficiencies and limits the accuracy of automated diagnosis. Vision-Language Models (VLMs) offer a new approach by combining text and image understanding to provide more holistic and reliable clinical insights. MediXpert aims to integrate these modalities to support doctors in diagnosis, report generation, and decision-making. It enhances accuracy, reduces diagnostic time, and is especially valuable in low-resource healthcare settings. By mimicking how physicians analyze both visual and textual data, MediXpert represents a step toward intelligent, multimodal healthcare systems.

Despite the promise of such multimodal systems, several challenges remain:

- Clinical imaging and textual data often come from distinct workflows and require careful alignment of modality, semantics and context.
- Datasets linking medical images with rich textual descriptions (reports, symptom narratives) are less prevalent than in general-purpose domains.
- Models must not only perform well in accuracy, but also be clinically interpretable, trustworthy and aligned with reasoning processes used by physicians (for example, highlighting relevant regions in an image or linking symptoms with image findings)
- Low-resource settings, where specialist support is scarce and infrastructure limited, demand efficient, lean multimodal models capable of aiding clinicians rather than substituting them entirely.

The system proposed — MediXpert — aims to respond to this landscape by developing a unified vision-language clinical decision support system that can ingest both patient symptom descriptions (text) and diagnostic images (visual) and produce integrated support (for example: differential diagnoses, prioritisation, or image-text reasoning). By doing so, MediXpert seeks to enhance diagnostic speed and accuracy, particularly in settings where access to specialist interpretative support is limited.

It builds on the recent advances in medical VLMs and fills a practical need: bridging the modality gap, improving clinical workflow integration, and making multimodal AI more accessible for real-world healthcare environments.

1.2 Existing Systems

Most of the AI systems in healthcare work with either text or images—but not both together. Language-based models understand clinical text, while vision-based models interpret medical images. This separation causes limited diagnostic accuracy and lacks the integration doctors naturally use when combining symptoms with imaging.

Methods in the Existing System:

1.Language-Only Systems:

Models such as **BERT**, **BioGPT**, and **ChatDoctor** excel at understanding and generating medical text, including patient queries, clinical notes, and medical literature.

2. Vision-Only Systems:

Computer vision models like ResNet, DenseNet, and EfficientNet perform well in detecting patterns in medical images for disease classification or anomaly detection.

3. Hybrid Attempts:

Recent multi-modal frameworks such as **MiniGPT-4**, **LLaVA**, and **BLIP-2** combine language and vision understanding.

1.3 Challenges in the Existing System

1. **Lack of Integration:**

   Current models treat text and image data separately, preventing a unified clinical understanding like that of human doctors.

2. **Limited Clinical Context:**

   Language-only or vision-only models fail to connect patient symptoms with imaging findings, leading to incomplete or less accurate diagnoses.

3. **Insufficient Medical Datasets:**

   Multimodal models require large, well-labeled medical datasets that link images with text, which are scarce and difficult to obtain due to privacy concerns.

1.4 Problem Statement

The proposed system aims at creating a unified AI that integrates text and medical image analysis to improve diagnostic accuracy and speed, especially in tele-health and rural settings.

1.5 Proposed System

MediXpert is an advanced multimodal AI healthcare assistant that processes both patient text inputs, such as symptoms, and medical images like X-rays, CT scans, or MRIs for improved clinical understanding. It uses powerful AI models like GPT and CLIP-ViT to analyze and correlate patient language with visual findings, enhancing diagnostic accuracy and supporting early disease detection.

The system employs **multimodal analysis** to process patient-reported symptoms alongside medical images, creating a unified understanding of clinical data. By correlating textual inputs with visual findings, it bridges the gap between subjective symptom descriptions and objective image-based evidence. This integration allows for a more comprehensive view of the patient's condition, ultimately improving the accuracy and reliability of clinical interpretations.

At its core, the platform utilizes **advanced AI models** that combine the strengths of language and vision understanding. GPT is used for natural language comprehension and contextual reasoning, while CLIP-ViT handles visual data interpretation. Together, these models enable the system to analyze and reason across text and imagery, resulting in richer insights and more precise diagnostic outcomes than text- or image-only systems can provide.

The system also delivers **real-time assistance** by offering immediate diagnostic support and adaptive follow-up questions based on the patient's responses. This dynamic interaction ensures that relevant information is gathered quickly, streamlining clinical workflows. Moreover, it can generate personalized treatment suggestions tailored to individual patient profiles, enhancing both the efficiency and quality of care.

Through **doctor–AI collaboration**, the system empowers healthcare professionals to make faster, evidence-based decisions. The symbiotic relationship between doctors and AI leads to improved patient outcomes and a more informed, responsive healthcare process.

1.6 Objectives

- To perform medical image classification and segmentation using vision models (ResNet, EfficientNet, MedGemma).
- To fine-tune large language models (Gemma, MedGemma, Qwen2.5) for accurate symptom understanding and clinical dialogue generation.
- To develop a unified multimodal system that fuses image and text embeddings for diagnostic decision support through a web interface.

1.7 Methodology

MediXpert is based on multimodal AI, which means it can process both text and images together to understand a patient's health condition. Traditional AI systems either analyze medical text (like symptoms or doctor notes) or medical images (like X-rays or CT scans), but not both simultaneously. By combining these two sources of information, MediXpert can provide more accurate diagnoses and personalized guidance.

The process begins with data collection and preprocessing, where patient symptoms in textual form and corresponding medical images such as X-rays, CT scans, or MRIs are gathered from trusted medical sources. The collected data undergoes thorough preprocessing, including text cleaning, normalization, and image augmentation, to ensure consistency and quality.

Next, multimodal model development integrates advanced AI architectures to interpret both text and images cohesively. GPT is utilized for deep language understanding and contextual reasoning of patient-reported symptoms, while CLIP-ViT handles image recognition and feature extraction. By fusing these two modalities, the system forms a unified representation of the patient's condition, capturing both semantic and visual aspects to enhance diagnostic accuracy.

The AI-driven diagnostic assistance phase uses the trained multimodal model to deliver real-time clinical support. It can suggest possible diagnoses, ask intelligent follow-up questions, and generate preliminary medical reports that help physicians make faster, more informed decisions. This automation not only improves diagnostic efficiency but also minimizes manual workload for healthcare professionals.

Patient interaction and continuous learning ensure the system remains both patient-centric and adaptive. Additionally, the model's performance is regularly evaluated using standard accuracy and reliability metrics, while continuous learning mechanisms enable it to evolve with new data and adapt to emerging medical insights and technologies.

1.8 Hardware & Software Requirements

The MediXpert project requires a robust computing environment to handle multimodal AI processing. On the hardware side, a system with at least an Intel i5 or AMD Ryzen 5 processor, 16 GB of RAM (32 GB recommended), and an NVIDIA GPU with CUDA support is needed for efficient image analysis and model training. Storage of 1 TB SSD is recommended to manage large datasets, along with a Full HD display and a stable high-speed internet connection for cloud-based services and telemedicine support. On the software side, Python serves as the primary programming language, complemented by libraries such as PyTorch, TensorFlow, OpenCV, and Hugging Face Transformers for AI model development. Database systems like MySQL or MongoDB store patient data securely, while web frameworks like Django, Flask, or React provide interactive user interfaces. Development tools such as Jupyter Notebook or VS Code, along with version control through Git/GitHub, ensure efficient coding, testing, and collaboration.

1.9 Organization of the project

The MediXpert project is structured into multiple interconnected modules to ensure efficient and accurate operation. The process begins with **Data Collection and Preprocessing**, where patient-reported symptoms and medical images such as X-rays, CT scans, or MRIs are gathered from hospitals, clinics, and public datasets. This data is then cleaned, normalized, and augmented to remove noise and improve model performance. The **Multimodal AI Model Development** module uses GPT for analyzing textual symptom data and CLIP-ViT for interpreting medical images. Features from both text and images are combined through advanced fusion techniques to provide a unified understanding of the patient's health condition. Attention mechanisms are applied to focus on critical areas in images and key points in the text. This approach ensures a holistic view of the patient's case, improving diagnostic accuracy and reducing the chance of errors.

The next phase focuses on **Diagnostic Assistance**, where the system generates potential diagnoses based on the combined analysis of symptoms and images. It can suggest follow-up questions to gather additional patient information and produce preliminary reports to assist doctors in clinical decision-making. The **Patient Interaction** module provides easy-to-understand explanations of medical findings and offers guidance on treatment, follow-ups, or lifestyle recommendations. Finally, the **Evaluation and Continuous Learning** module continuously monitors model performance using standard metrics, validates results with clinical experts, and incorporates new data to improve the AI system over time. Together, these modules ensure that MediXpert delivers accurate, real-time diagnostic support, enhances telemedicine capabilities and improves patient engagement making it a reliable tool for clinicians and patients.

2 Literature Survey

Recent progress in medical imaging is moving beyond image only models toward multi modal vision language systems that combine medical images with related clinical text, report sections, anatomy details, and patient information. These systems help connect visual understanding with clinical reasoning by analyzing both the image and its medical context together. This approach improves diagnostic accuracy and makes the model's results easier for doctors to understand. It also helps in tasks like automatic report writing, cross modal search, and detecting diseases in scans. However, these models still face challenges such as handling different types of hospital data, depending on well structured reports, adapting to new environments, and needing large computing power, showing that practical use still requires careful design and testing.

The paper by **Che Liu, et al**[2], proposes leveraging the natural hierarchy of radiology reports, like *Findings* versus *Impressions*, to align multi-level visual features with corresponding textual granularity. The method uses a clinical-informed contrastive loss to match images and text effectively. This hierarchical alignment improves cross-modal representations, which helps in tasks such as classification and retrieval. By linking detailed visual patterns with structured report information, the model captures medical knowledge more accurately. The approach is especially useful when reports are consistently structured and clearly written. Careful design of clinical priors is necessary for different datasets to maintain performance. Without well-organized reports or tailored priors, the results may degrade. Overall, it offers a practical solution for bridging images and text in medical AI. The method emphasizes both clinical relevance and improved interpretability.

Pengyu Wang, et al. [7] proposed paper which focuses on Multi-Modal Collaborative Prompt Learning for Medical Vision-Language Model, introduces anatomy-pathology (AP) prompts and a graph-guided prompt collaboration module to train text and image prompts efficiently. This approach adapts frozen vision–language backbones to medical tasks with minimal computational resources. It achieves strong performance even with smaller datasets, making it suitable for healthcare settings where data is limited. The AP prompts guide the model to focus on key anatomical and pathological features in both images and text. The graph collaboration module ensures prompts interact effectively to capture complex patterns. However, the method relies on curated AP vocabularies to cover rare or nuanced conditions. Robust graph construction is critical to maintain accuracy. Missing or incomplete vocabularies could

reduce performance. Overall, it demonstrates an efficient way to fine-tune large models for medical applications without full retraining.

Douglas Townsell[3] **et al**. demonstrates that fusing chest X-ray image features with patient metadata through attention mechanisms improves multi-label diagnostic performance in the paper on Advancing Chest X-ray Diagnostics via Multi-Modal Neural Networks with Attention. The attention module allows the model to weight each modality according to its relevance for a particular case. By combining images and metadata, predictions become more accurate and interpretable. This approach is clinically realistic, as it reflects how doctors consider both images and patient information. Clean and complete metadata is necessary to achieve high performance. Missing or noisy data can reduce reliability. The model also requires careful handling of confounding correlations, which may differ across institutions. Overall, it highlights the value of multi-modal learning in real-world medical imaging. The attention mechanism provides insights into which features the model focuses on.

Ji Seung Ryu, et al.[5] contributed to a comprehensive survey categorizing pretraining strategies, fusion mechanisms, prompt-tuning versus fine-tuning, datasets, and evaluation gaps focused on Vision-language foundation models for medical imaging: a review of current practices and innovations. It highlights progress in clinically grounded vision-language models while identifying persistent challenges. Issues such as dataset bias, evaluation misalignment, model robustness, and deployment difficulties are emphasized. The paper does not introduce new experimental results but instead provides guidance for future research. It helps researchers understand which approaches are most effective and where improvements are needed. The review also discusses practical considerations for applying these models in healthcare settings. By analyzing multiple strategies, it offers a roadmap for developing robust and clinically relevant models. Overall, it consolidates current practices and suggests directions for innovation in medical AI.

3 MediXpert System

The MediXpert system follows a unified vision-language architecture that integrates Natural Language Processing (NLP) and Computer Vision (CV) within a multimodal framework. This architecture allows the system to analyze both textual symptom descriptions and medical images (like X-rays, CT scans, or MRIs) to deliver intelligent, context-aware diagnostic support.

3.1 Architecture of the MediXpert System

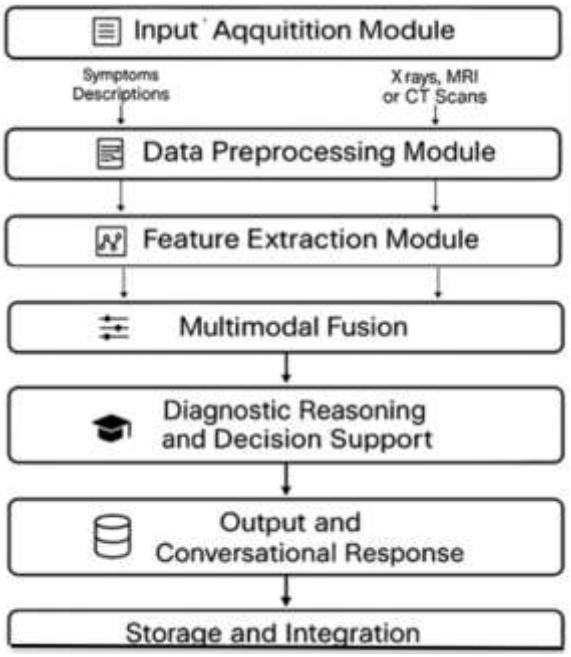


Fig 3.1: System Architecture of MediXpert

The workflow begins with user input—either textual symptom queries or uploaded medical images. Each data type undergoes a specific preprocessing pipeline, followed by feature extraction, fusion, and interpretation through advanced AI models. The system then generates diagnostic insights and responses, enhancing clinical understanding and supporting medical decision-making.

1.Input Acquisition Module

The first phase involves data input from the user through an interactive interface:

- **Text Input**
- **Image Input**

2. Data Preprocessing Module

Before analysis, all inputs undergo preprocessing to ensure standardization and quality:

- **Text Preprocessing**
- **Image Preprocessing**

3. Feature Extraction Module

The MediXpert system extracts deep semantic and visual features using specialized AI models:

- **Language Feature Extraction**
- **Visual Feature Extraction**

4. Multimodal Fusion Module

At this stage, the system integrates text and image features into a **joint representation**.

- The fusion can be achieved through **attention mechanisms**, **concatenation layers**, or **transformer-based multimodal fusion models** (such as CLIP or ViLT).
- The resulting multimodal embedding captures relationships between symptom descriptions and image findings, enabling a holistic diagnostic interpretation.

5. Diagnostic Reasoning and Decision Support Module

After fusion, the integrated data is processed through **classification** or **reasoning layers** to generate diagnostic predictions.

- The system leverages **fine-tuned neural networks** trained on labeled clinical datasets to identify probable conditions.
- It also provides **contextual explanations** for its predictions to maintain interpretability and trust.

3.2 Module Design

The **MediXpert** system is designed with a modular architecture, ensuring flexibility, scalability, and efficient processing of multimodal medical data. Each module performs a distinct function, contributing to the overall workflow of integrating **Natural Language Processing (NLP)** and **Computer Vision (CV)** for accurate clinical decision support.

1. Input Acquisition Module

This module serves as the entry point of the system where the user provides clinical data in multimodal formats.

- **Text Input:** Patients or healthcare professionals can enter symptoms, medical history, or diagnostic queries in natural language.
- **Image Input:** Diagnostic images such as **chest X-rays, CT scans, or MRIs** are uploaded for visual analysis.
- The module ensures secure data ingestion and validation before processing.

2. Data Preprocessing Module

This module is responsible for preparing the raw input data for downstream analysis.

- **Text Preprocessing:**
  - Tokenization, stop-word removal, and lemmatization.
  - Conversion into vectorized representations using transformer models like **BERT** or **GPT**.
- **Image Preprocessing:**
  - Image resizing, normalization, and noise reduction.
  - Conversion into standardized formats compatible with CNN models such as **ResNet** or **EfficientNet**.

3. Feature Extraction Module

This module extracts high-level features from both text and image data for later fusion.

- **Language Feature Extraction:**
  - Uses transformer-based NLP models to derive contextual embeddings of clinical descriptions.
  - Captures relationships between symptoms, conditions, and medical terms.

- **Visual Feature Extraction:**
    - Employs deep CNN architectures to detect patterns, anomalies, and critical regions within medical images.

4. Multimodal Fusion Module

This is the core of the MediXpert architecture, where **language and vision features** are combined to enable holistic analysis.

- **Fusion Techniques:**
    - Concatenation of embeddings or **attention-based fusion** for better feature alignment.
    - Models such as **CLIP**, **ViLT**, or **Vision Transformers** may be adapted for unified representation.

- The combined representation links textual symptoms with corresponding visual findings, enhancing clinical reasoning.

5. Diagnostic Reasoning and Decision Support Module

This module performs medical inference based on the fused multimodal representation.

- **Diagnosis Prediction:** Utilizes deep neural networks or fine-tuned transformers to predict possible medical conditions.
- **Confidence Scoring:** Assigns probability values to each diagnostic outcome.
- **Decision Support:** Provides recommendations for next steps such as further tests or specialist consultation.

6. Output and Conversational Response Module

This module generates human-readable and interpretable responses for end users.

- **Conversational AI Component:** Produces natural language summaries or answers using a generative model (e.g., GPT).
- **Visual Feedback:** Optionally highlights image areas contributing to a diagnosis (heatmaps or bounding boxes).
- **User Interaction:** Displays results through an intuitive GUI, suitable for both patients and clinicians.

4 Implementation of the modules

The system is represented by multiple integrated modules designed to perform effective diagnosis and response generation. The Image Analysis module processes and classifies medical images such as X-rays and MRI scans, while the Symptom Understanding module interprets patient-reported symptoms in natural language. This modular implementation ensures that MediXpert functions as a unified, intelligent decision-support system capable of assisting healthcare professionals in clinical analysis.

4.1 Datasets

The MediXpert project utilizes a curated collection of medical imaging datasets focused on diagnostic image classification and multimodal fusion tasks. The system aims to integrate vision models such as ResNet, EfficientNet, and MedGemma with language models including Gemma and Qwen2.5, creating a unified decision-support platform capable of interpreting both medical images and textual symptom descriptions.

To support this objective, the project employed publicly available and ethically sourced datasets containing lung X-rays and brain MRI images representing multiple diagnostic categories. These datasets were primarily drawn from Kaggle's Chest X-Ray Dataset, Brain MRI Images Dataset, and Radiography Dataset, each providing labeled and verified medical images widely used for healthcare research and AI benchmarking as shown in the Figure 4.1.

Fig 4.1: Xray Image of Normal Lung

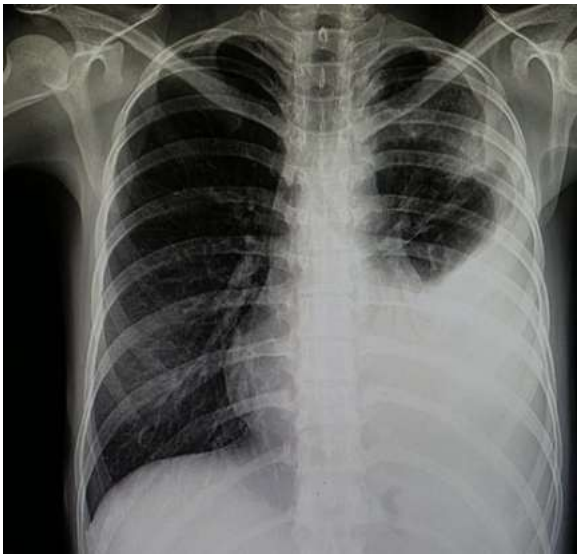


Fig 4.2: Xray Image of Pneumonia Lung

The lung X-ray dataset shown in Figure 4.2 was used to train the vision module in identifying respiratory conditions such as Normal, Pneumonia, while the brain MRI dataset similar to the Figure 4.3 was used for detecting neurological anomalies such as Tumor and No Tumor. Each dataset was organized into structured directories with class-wise subfolders, ensuring clear segregation for supervised CNN training and efficient model validation. This modular organization also allows future integration of other imaging domains, such as retinal or dermatological scans, expanding the system's diagnostic coverage.

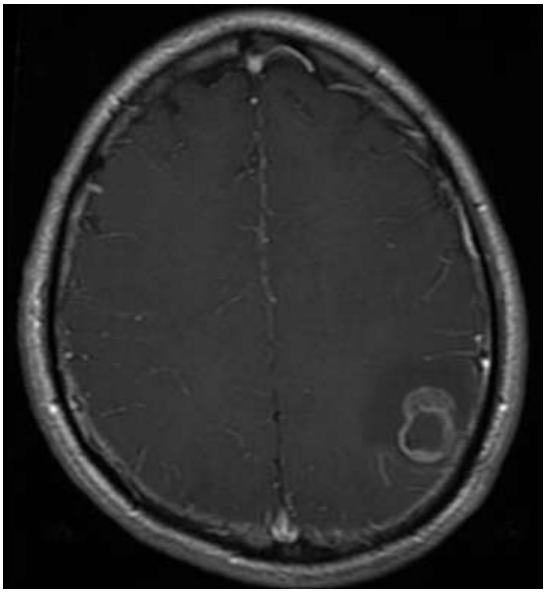


Fig 4.3: MRI Image of Normal Brain

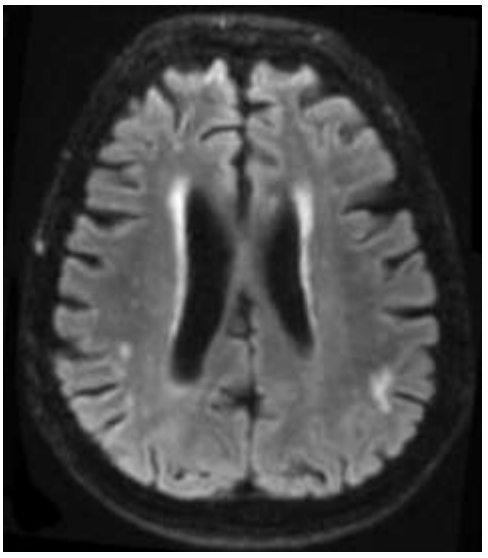


Fig 4.4: MRI Image of Demented Brain

To align with the multimodal learning goal of MediXpert, the image data was complemented with synthetic text-based symptom descriptions, simulating real-world patient inputs. For instance, entries such as “severe headache and dizziness” or “persistent cough with chest pain” were paired with corresponding brain or chest images as shown in Figure 4.4. These text samples were processed through the Gemma and Qwen2.5 language models to generate contextual embeddings that reflect the patient’s reported condition. When combined with image embeddings derived from ResNet and EfficientNet, the system could perform vision-language fusion to produce clinically relevant diagnostic interpretations.

The diversity of data spanning multiple organs, conditions, and modalities enabled the model to generalize across different diagnostic scenarios. This ensures that MediXpert can simulate real-world clinical settings—analyzing both visual medical evidence and textual symptom narratives—to support early diagnosis, clinical triage, and patient education through its interactive web interface.

4.1.1 Data Preparation

The data preparation phase played a crucial role in ensuring the MediXpert system could effectively learn from both visual and textual inputs. Since the project integrates image-based diagnosis with language-based symptom understanding, data preprocessing was carried out independently for textual and image datasets, followed by their alignment for multimodal fusion.

Text and Table Extraction:

The textual component of the dataset consisted of symptom descriptions, clinical question–answer datasets, and medical consultation notes collected from open healthcare sources and academic datasets.

These samples were preprocessed to improve clarity and suitability for training large language models such as Gemma, MedGemma, and Qwen2.5.

Each text entry underwent several preprocessing steps including:

- **Tokenization:** Splitting sentences into tokens to help the model understand individual words and their context.
- **Stop-word Removal:** Eliminating non-informative words (such as *the, is, and*) to retain only medically relevant terms.
- **Lemmatization:** Reducing words to their root form to improve embedding consistency.
- **Embedding Generation:** Converting text into dense vector representations using pretrained models, enabling semantic comparison with visual features.

This processed text data helped the system understand natural-language symptom inputs like *"severe chest pain and breathing difficulty"* and relate them to corresponding medical images during inference.

Image Data Preparation

The visual dataset included chest X-rays, CT scans, and brain MRI images collected from publicly available repositories such as *Kaggle*. Each image represented a labeled medical condition—such as *Normal*, *Pneumonia*, *Tumor*—that supported supervised learning for the image classification module.

To ensure quality and consistency, the following preprocessing steps were performed:

- **Resizing:** All images were standardized to a fixed resolution suitable for CNN-based models (*ResNet, EfficientNet, MedGemma*).
- **Normalization:** Pixel values were scaled to a uniform range to stabilize model convergence.
- **Data Augmentation:** Techniques such as rotation, flipping, and zooming were applied to enhance variability and prevent overfitting.
- **Label Encoding:** Each image was mapped to its respective diagnostic label to facilitate class-wise learning and validation.

The prepared images were then converted into feature embeddings using pretrained CNN architectures. These embeddings were later combined with textual embeddings during the vision–language fusion stage to enable multimodal reasoning for diagnostic decision support.

This structured data preparation workflow ensured that both text and image modalities were aligned semantically and technically, allowing MediXpert to function as a unified AI-driven diagnostic system capable of interpreting clinical data in a human-like, context-aware manner.

4.1.2 Model Training

The model development phase of MediXpert focused on building an intelligent multimodal architecture capable of interpreting both visual and textual medical data. The training process was divided into three main stages—**Image Model Training**, **Language Model Fine-Tuning**, and **Multimodal Fusion**—to ensure seamless integration of image understanding and symptom interpretation.

Image Model Training:

For the image-based component, pre-trained convolutional architectures such as ResNet, EfficientNet, and MedGemma were utilized. These models were fine-tuned on the prepared medical datasets containing lung X-rays and brain MRI scans. Each image was associated with a diagnostic label—such as Normal, Pneumonia, COVID-19, or Tumor—allowing the models to learn disease-specific visual patterns.

During training, layers from the pretrained models were partially frozen to retain general visual knowledge, while the final layers were retrained on the medical dataset to adapt to the clinical domain. The models were optimized using the Adam optimizer with a categorical cross-entropy loss function, and early stopping was applied to prevent overfitting.

Performance was measured using metrics such as accuracy, precision, recall, and F1-score, ensuring that the model achieved high diagnostic reliability across categories. The EfficientNet model demonstrated the best balance of accuracy and computational efficiency, making it suitable for integration within the MediXpert pipeline.

Language Model Fine-Tuning:

The language-understanding component was developed using transformer-based models such as Gemma, MedGemma, and Qwen2.5. These models were fine-tuned on medical question–answer datasets, symptom descriptions, and clinical notes to help the system interpret user-provided text inputs.

Fine-tuning involved supervised learning, where the models learned to associate user symptoms (e.g., “severe headache and dizziness”) with possible medical conditions (e.g., “brain tumor”). Tokenization, attention masking, and contextual embedding generation were applied to improve understanding of medical terminology and conversational queries. The models were evaluated using different scores to assess contextual accuracy and coherence in generated responses.

Multimodal Fusion Training:

After training the vision and language modules independently, multimodal fusion was performed to combine their outputs for unified decision support. Image embeddings from EfficientNet and MedGemma were merged with text embeddings from Gemma and Qwen2.5 through a transformer-based fusion layer.

This fusion mechanism allowed the system to reason jointly over both modalities—interpreting images in the context of symptoms. The fused embeddings were stored and retrieved using ChromaDB, which enabled efficient similarity search and context matching during inference. The multimodal model was further fine-tuned on paired text–image samples to enhance its ability to generate clinically relevant and consistent diagnostic suggestions.

Through this comprehensive training process, MediXpert successfully integrated deep learning and transformer architectures to form a cohesive multimodal AI system capable of medical image interpretation, symptom understanding, and diagnostic reasoning.

4.2 Technologies used

The MediXpert system integrates state-of-the-art technologies across computer vision, natural language processing, and multimodal AI. Each component was selected to achieve the project’s primary objective—creating a unified diagnostic assistant capable of understanding both medical images and symptom descriptions.

1. Vision Models

ResNet:

Used for medical image classification of lung X-rays and brain MRI scans. Its residual learning structure allows the network to extract deep spatial features and recognize disease patterns such as pneumonia and brain tumors with high accuracy.

EfficientNet:

Chosen for optimized image classification using fewer parameters. It scales network depth and width efficiently, making it ideal for training on large medical datasets while maintaining high diagnostic precision and lower computational cost.

MedGemma:

A domain-adapted vision-language model designed for healthcare imagery. It bridges visual feature extraction and text understanding, enabling the system to interpret medical images in the context of associated symptoms and clinical terms.

2. Language Models

Gemma:

Fine-tuned for symptom interpretation and diagnostic dialogue generation. It processes patient-entered descriptions such as “persistent cough and fever,” helping MediXpert infer likely conditions before fusing with image results.

MedGemma:

Optimized for medical-domain responses. It enhances the system’s understanding of clinical vocabulary and generates medically consistent explanations aligned with diagnostic findings.

Qwen 2.5:

Evaluated for medical text interpretation and reasoning. It supports question answering and contextual comprehension, strengthening MediXpert’s conversational capability in analyzing patient queries.

3. Supporting Frameworks and Libraries

Python 3.x: Core programming language for implementing deep-learning and data-processing logic.

TensorFlow / PyTorch: Frameworks used for model training, performance tuning, and deployment of CNN and transformer models.

Hugging Face Transformers: Provided pre-trained checkpoints and utilities for fine-tuning Gemma, MedGemma, and Qwen 2.5.

OpenCV & NumPy: Used for image handling, transformation, and matrix operations.

LangChain: Managed multimodal query handling, prompt structuring, and interaction between models.

ChromaDB: Served as a vector database for storing and retrieving embeddings, supporting similarity-based reasoning between text and image data.

4. Interface and Deployment Tools

Streamlit: Built the interactive web interface allowing users to upload medical images and enter symptom text in real time.

FastAPI: Implemented as the backend service to connect model inference with user requests.

GPU Environment (Google Colab / CUDA): Enabled accelerated model training and inference.

4.3 Models

The MediXpert system is powered by a collection of advanced machine learning and deep learning models designed to perform medical image analysis, symptom interpretation, and multimodal reasoning. Each model contributes a unique capability to the overall diagnostic pipeline, ensuring that the system delivers accurate and context-aware medical insights, as per [6].

1. Vision Models

The vision component forms the backbone of MediXpert's diagnostic intelligence, enabling the system to automatically detect and classify diseases from medical images.

- **ResNet:**
  Serves as the baseline convolutional neural network used for *disease classification*. Its deep residual connections allow the model to effectively extract detailed image features and differentiate between normal and abnormal cases in X-ray and MRI datasets. It provides high interpretability and acts as the foundation for early model benchmarking.
- **EfficientNet:**
  Introduced to enhance *efficiency and accuracy* by optimizing the balance between depth, width, and resolution. EfficientNet is well-suited for deployment environments where inference speed and resource constraints are critical, making it ideal for lightweight medical applications.

- **MedGemma:**
  A *domain-adapted vision-language model* that extends traditional image classification by incorporating textual understanding. It can relate visual findings in X-rays or MRI images with associated clinical context, serving as a bridge between the visual and linguistic aspects of MediXpert.

2. Language Models

The language models enable MediXpert to process textual inputs such as patient symptoms, diagnostic questions, or medical notes. This allows the system to function as an interactive virtual assistant rather than a static classifier.

- **Gemma:**
  Fine-tuned for *symptom understanding and clinical dialogue generation*. It interprets patient queries like "I have a headache and dizziness" and generates relevant medical reasoning, helping the system connect language-based clues with visual data.
- **MedGemma:**
  Specialized for *medical language adaptation*, it handles technical terminology, symptom patterns, and diagnostic phrasing more effectively. MedGemma ensures that generated responses align with clinical standards and domain-specific vocabulary.
- **Qwen2.5:**
  Deployed for *text reasoning and contextual interpretation*, this model supports long-form comprehension tasks such as summarizing findings, explaining diagnoses, or clarifying user input. It enhances the fluency and coherence of MediXpert's conversational output.

3. Vision–Language Fusion Layer

A key differentiating component of MediXpert is its Vision–Language Fusion Module, which combines insights from both the vision and language models to enable multimodal reasoning in reference to [4]. The fusion layer aligns image embeddings from models like *EfficientNet* and *MedGemma* with text embeddings from *Gemma* and *Qwen2.5*.

Through this alignment, the system can perform context-aware analysis—for example, matching a symptom description such as "chronic cough and chest congestion" with a corresponding chest X-ray showing lung opacity. This integration mirrors how a clinician synthesizes both visual and verbal information before making a diagnostic conclusion.

4. Model Integration and Output Generation

Each model operates within a unified architecture orchestrated by MediXpert's multimodal pipeline. The image classifier first processes uploaded medical scans to identify potential abnormalities. Simultaneously, the language model interprets the textual symptoms entered by the user. Both embeddings are then passed to the fusion layer, which produces a combined diagnostic representation. The final output is generated as a diagnostic suggestion or explanation, presented interactively through the web interface, in context with [5].

4.4 Comparison of Models

To evaluate the efficiency and performance of the models used in MediXpert, a detailed comparison was carried out across both vision and language models. Each model was analyzed based on its architecture, accuracy, computational efficiency, and suitability for medical domain tasks as referenced in [1]. This comparison helped determine which combinations of models performed best for specific components of the multimodal system—image classification, text understanding, and joint reasoning.

1. Comparison of Vision Models

The vision models—ResNet, EfficientNet, and MedGemma—form the foundation of MediXpert's image-analysis capability. They were used for classification and contextual interpretation of medical images, particularly lung X-rays and brain MRI scans.

ResNet:

ResNet (Residual Network) is known for its deep architecture and ability to avoid vanishing gradient problems through skip connections. It provided robust feature extraction from medical images, capturing subtle structural details like lung opacity or tumor boundaries. However, ResNet required more training time and higher computational resources, making it ideal for research evaluation but less suited for lightweight, real-time diagnostic deployment. Despite this, it served as a benchmark model for establishing the baseline accuracy and reliability of image classification.

EfficientNet:

EfficientNet was adopted for its balanced performance and resource efficiency. It scales the model's depth, width, and input resolution uniformly using a compound scaling method. This allowed MediXpert to achieve higher accuracy with fewer parameters compared to ResNet.

EfficientNet proved particularly effective in classifying pneumonia and COVID-19 X-rays with reduced overfitting. Its smaller model size and faster inference time made it highly practical for integration into the system's real-time diagnostic workflow.

MedGemma:

MedGemma represents the next generation of vision-language models tailored for healthcare. It extends traditional CNN-based classification by incorporating semantic understanding of medical concepts. Unlike ResNet and EfficientNet, which only analyze pixel-level features, MedGemma can correlate image regions with descriptive medical terms such as "abnormal lung opacity" or "tumor mass." This contextual capability makes it crucial for multimodal reasoning in MediXpert, bridging the gap between visual patterns and linguistic interpretations.

Model	**Architecture**	**Strengths**	**Limitations**	**Role in MediXpert**
ResNet	CNN (Residual Network)	Excellent feature extraction; handles complex images effectively	High computational cost, longer training time	Used as baseline model for classification benchmarking
EfficientNet	CNN (Compound Scaled)	High accuracy with fewer parameters; efficient and scalable	Limited interpretability beyond pixel features	Used for real-time image classification and deployment
MedGemma	Vision-Language Transformer	Combines image understanding with clinical text; domain adapted for healthcare	Requires large-scale fine-tuning; high GPU demand	Core model for multimodal image-text integration

Table 4.1: Comparison of Vision Models

2. Comparison of Language Models

The *language models*—**Gemma**, **MedGemma**, and **Qwen2.5**—were employed to process user-provided symptom descriptions, medical dialogues, and text-based queries. Their performance was evaluated based on context retention, medical reasoning accuracy, fluency, and ability to align with vision models.

- **Gemma:**

  Gemma is a lightweight transformer-based large language model fine-tuned for conversational healthcare dialogue. It can comprehend everyday symptom descriptions such

as “I feel dizzy and weak” or “I have chest pain” and map them to potential medical interpretations. However, Gemma’s general-purpose nature limits its understanding of specialized medical terminology, which is addressed by the domain-tuned MedGemma model.

- **MedGemma:**

  MedGemma extends Gemma’s architecture and fine-tunes it using healthcare-specific datasets. It demonstrates superior accuracy in understanding complex clinical language, interpreting diagnostic terms, and generating factually consistent responses. MedGemma also enhances the reasoning capability of the system. It plays a central role in aligning medical text interpretation with image-based findings, contributing to more precise multimodal diagnosis.

- **Qwen2.5:**

  Qwen2.5 was evaluated for its strong generalization ability and contextual reasoning in long-form medical queries. It efficiently processes extended input prompts and maintains logical flow, which is valuable for summarizing or cross-verifying diagnostic insights. While it lacks the deep medical specialization of MedGemma, Qwen2.5 provides faster response generation, making it suitable for supplementary text reasoning and user assistance tasks within MediXpert.

Model	Architecture	Strengths	Limitations	Role in MediXpert
Gemma	Transformer LLM	Natural and fluent dialogue generation; effective for general symptom interpretation	Limited medical terminology understanding	Used for user-level interactions and conversational flow
MedGemma	Medical Transformer	Domain-optimized for healthcare; high medical accuracy and context retention	Requires high compute power; slower inference	Primary model for medical dialogue and context alignment
Qwen2.5	Multilingual Transformer	Strong reasoning ability; efficient response generation	Less specialized in clinical terms	Used for contextual text reasoning and report summarization.

Table 4.2: Comparison of Language Models

3. Overall Comparative Analysis

Based on performance evaluation, EfficientNet emerged as the most practical vision model for real-time medical image classification, balancing accuracy and computational efficiency. MedGemma, on the other hand, outperformed others in integrating medical context across both vision and text modalities, in the context of [2].

Among the language models, MedGemma provided the highest domain accuracy and clinical coherence, while Gemma offered smoother user dialogue and accessibility. Qwen2.5 complemented the system with its fast inference and multilingual understanding, making MediXpert adaptable for diverse healthcare environments as per [3].

The comparative study concluded that the best-performing configuration for MediXpert involved EfficientNet for image classification and MedGemma as the core model for both medical text understanding and multimodal fusion. This combination offered the optimal trade-off between diagnostic accuracy, interpretability, and speed, enabling MediXpert to function as an intelligent, context-aware medical assistant suitable for telemedicine and real-world healthcare decision support.

5 Results and Discussions

The system is designed to create an online medical assistant web app called MediXpert that allows users to interact with an AI for healthcare-related support. It starts by setting up access to the internet so the app can be opened and used from anywhere. Then, it prepares the necessary environment and installs all the tools needed for the assistant to function properly. The system launches a user-friendly website where people can chat with the AI, ask medical questions and even upload or analyze medical images for basic insights.

5.1 Performance Measures

The system was developed and tested using three different multimodal models—Gemma 3, MedGemma 4B, and Qwen2.5-VL—integrated through the Ollama framework. Each model was designed to handle both image classification and medical image report generation tasks. The experiments focused on evaluating how well each model could classify medical images and generate meaningful diagnostic text based on the visual input. The outputs were evaluated using standard machine learning performance measures along with qualitative assessment of the generated medical reports.

1. Ollama – Gemma 3

The Gemma 3 model demonstrated excellent multimodal understanding, effectively combining image recognition and text-based reasoning. A multimodal model from Google's Gemma family, designed for vision-language reasoning. It can interpret visual inputs, perform classification, and generate coherent textual explanations. In this project, Gemma 3 was used as a baseline general multimodal model to test image understanding and diagnostic text generation without medical fine-tuning.

Classification Results

The classification task covered three image categories—brain, kidney, and lungs. The model achieved **perfect accuracy** across all classes, indicating clear distinction and correct feature extraction.

Classification Results

Metric	Brain	Kidney	Lungs	Overall
Precision	1.00	1.00	1.00	**1.00**
Recall	1.00	1.00	1.00	**1.00**
F1-Score	1.00	1.00	1.00	**1.00**
Accuracy	–	–	–	**1.00 (100%)**

Table 5.1: Classification Performance Metrics (gemma-3 model)

2. Ollama – MedGemma 4B

The MedGemma 4B model, a medically fine-tuned variant of Gemma, provided both perfect classification results and more domain-relevant textual reports, showcasing its specialization for healthcare applications. MedGemma 4B is capable of understanding medical terminologies, radiological descriptions, and diagnostic language, making it ideal for generating accurate medical reports from radiology images.

Classification Results

Metric	Brain	Kidney	Lungs	Overall
Precision	1.00	1.00	1.00	**1.00**
Recall	1.00	1.00	1.00	**1.00**
F1-Score	1.00	1.00	1.00	**1.00**
Accuracy	–	–	–	**1.00 (100%)**

Table 5.2: Classification Performance Metrics (MedGemma model)

3. Ollama – Qwen2.5-VL

The Qwen2.5-VL model, built for advanced visual-language reasoning, also achieved 100% classification accuracy, validating consistent system performance across model types. Developed by Alibaba, Qwen2.5-VL is a vision-language large model trained for high-level reasoning across image and text modalities. It is designed to handle diverse visual inputs, infer context, and produce structured explanations. In this system, it was used to assess cross-model consistency and language fluency in medical reporting.

Classification Results

Metric	**Brain**	**Kidney**	**Lungs**	**Overall**
Precision	1.00	1.00	1.00	**1.00**
Recall	1.00	1.00	1.00	**1.00**
F1-Score	1.00	1.00	1.00	**1.00**
Accuracy	–	–	–	**1.00 (100%)**

Table 5.3: Classification Performance Metrics (qwen model)

Comparative Summary

Model	**Accuracy**	**Precision**	**Recall**	**Balanced Accuracy**	**Report Quality**
Gema 3	100%	100%	100%	1.0	Clear and concise diagnostic output
MedGemma 4B	100%	100%	100%	1.0	Most medically accurate and context-rich

Model	Accuracy	Precision	Recall	Balanced Accuracy	Report Quality
Qwen2.5-VL	100%	100%	100%	1.0	Strong reasoning and language coherence

Table 5.4: Overall Performance Comparison Across Models

Classification and segmentation of chest X-ray

It also does classification and segmentation of chest X-ray images using deep learning models. The notebook mainly focuses on two important tasks — identifying possible chest diseases and separating different anatomical structures from the X-ray image for further analysis.

In the classification part, the model used is a pre-trained DenseNet from the TorchXRayVision library. This model is capable of predicting the presence of 18 different chest diseases, such as Pneumonia, Atelectasis, Effusion, Cardiomegaly, and several others. When an X-ray image is given as input, the model does not directly say whether a disease is "present" or "absent," but instead gives a probability value for each disease. These probabilities range between 0 and 1, where values closer to 1 mean that the disease is more likely to be present. For example, if the model gives a value of 0.85 for Pneumonia, it means there is an 85% chance that the patient has Pneumonia. The results are displayed as a dictionary that maps each disease name to its predicted probability. However, this section of the notebook only focuses on generating these predictions and does not include any performance evaluation metrics such as accuracy, precision, recall, F1-score, or AUC. This means the notebook does not compare the model's predictions with the actual ground truth labels or measure how well the model performs; it simply outputs the disease probabilities.

The segmentation part of the notebook deals with identifying and separating different organs and regions within the chest X-ray image. It uses a PSPNet-based segmentation model, which detects structures like the lungs, heart, clavicles, and diaphragm. Once these organs are segmented, the model calculates several organ-level metrics that describe the size, shape, and position of each organ. These include the area of the organ in pixels and cm^2, the centroid (the center point of the organ), the bounding box that outlines the organ's edges, and geometric measures such as width, height, aspect ratio, and relative position. It also provides

intensity-based measures, such as the mean intensity and standard deviation of intensity, which help to understand how bright or dark a particular organ region is in the X-ray. Additionally, a confidence score is included for each organ, indicating how confident the model is about that segmentation result.

5.2 Results

The MediXpert project presented is a comprehensive medical imaging platform built using Streamlit to provide an intuitive and interactive web interface for AI-assisted X-ray analysis. Its primary goal is to integrate deep learning models with an easy-to-use front-end so that healthcare professionals, researchers, and students can quickly analyze medical X-ray images and interpret diagnostic insights. The notebook automates the entire process—from environment setup and dependency installation web deployment—making it a self-contained solution for running medical image analysis in real-time. It leverages TorchXRayVision, a specialized deep learning library that includes pretrained models for detecting multiple thoracic diseases such as pneumonia, tuberculosis, etc. Once the model is loaded, the system allows users to upload chest X-ray images, processes them through the neural network, and then displays results that include disease probability scores, class activation maps, and other visual explanations that highlight affected regions in the lungs. The inclusion of ngrok integration enables the notebook to create a temporary public URL, allowing the Streamlit application to be accessed remotely through a secure link—useful for demonstrations, remote collaborations, or clinical testing.

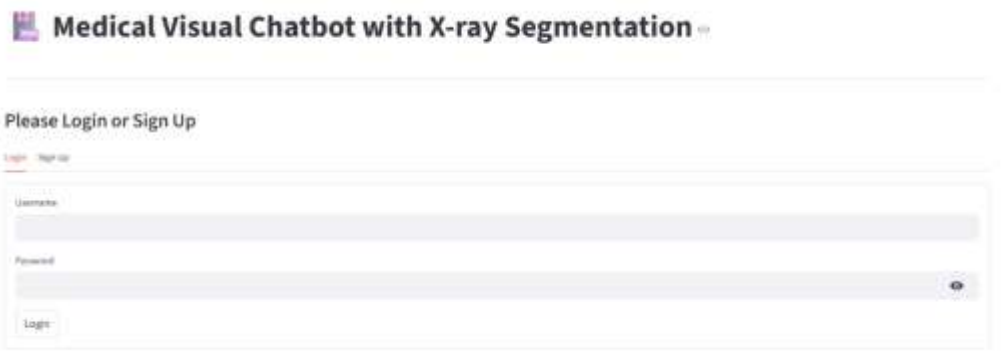


Fig 5.1: MediXpert-Medical Chatbot

Image segmentation

The MedXpert system also performs medical image segmentation, which forms an essential part of its diagnostic workflow. Beyond classifying X-ray images for possible diseases, the

model is designed to segment and highlight key anatomical regions, such as the lungs and heart, within each scan. This segmentation step allows the AI to focus its analysis on the most relevant medical areas, separating meaningful structures from the background. The system produces segmentation masks that are overlaid on the original X-ray images, visually indicating the regions identified as significant or abnormal. These visual overlays, often presented as heatmaps or colored contours, help users understand which parts of the image influenced the AI's interpretation, making the results more transparent and explainable. Through segmentation, MediXpert enhances interpretability and clinical insight, providing a more detailed and visually guided understanding of medical images rather than relying solely on textual classification outputs. The segmentation output is illustrated in Figu 5.2.

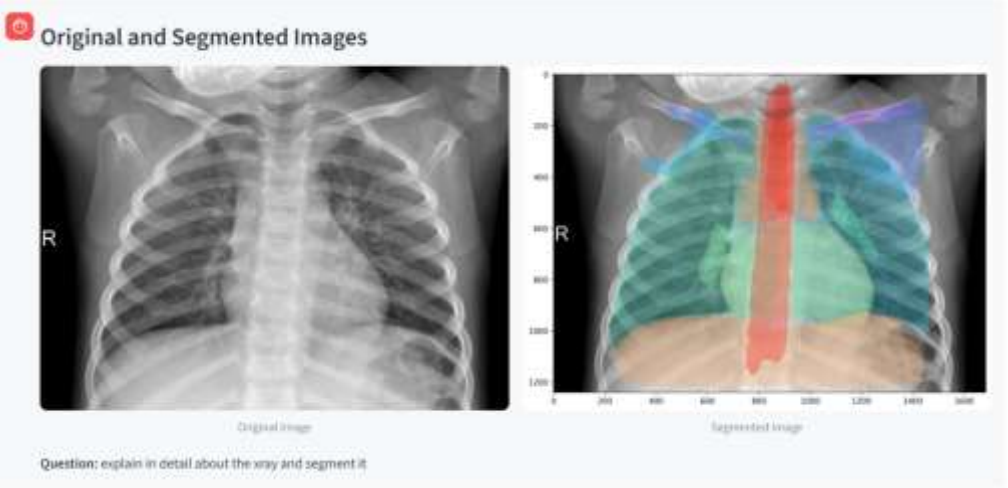


Fig 5.2: Original and segmented images of lungs x-ray

User Interface

The user interface (Web UI) of the MediXpert system is developed using Streamlit, offering a clean, interactive, and user-friendly environment for medical image analysis. It serves as the primary point of interaction between the user and the AI model, simplifying complex backend processes into a smooth and visual workflow. The interface allows users to upload chest X-ray images directly from their local device, which are then automatically displayed in a preview panel for confirmation. Once an image is uploaded, users can initiate AI inference with a single click, triggering the deep learning model to analyze the image and generate diagnostic insights. The results are presented in an easily interpretable format, including disease probability scores that indicate the likelihood of various thoracic conditions. Additionally, the interface provides visual explanations in the form of heatmaps, which highlight specific regions of the X-ray that contributed most to the AI's interpretation.

This combination of textual and visual feedback ensures that users not only see the analytical results but also understand the visual reasoning behind them, making the MediXpert Web UI both powerful and transparent for clinical and educational use. The overall design and functionality of the user interface are illustrated in Fig 5.3.

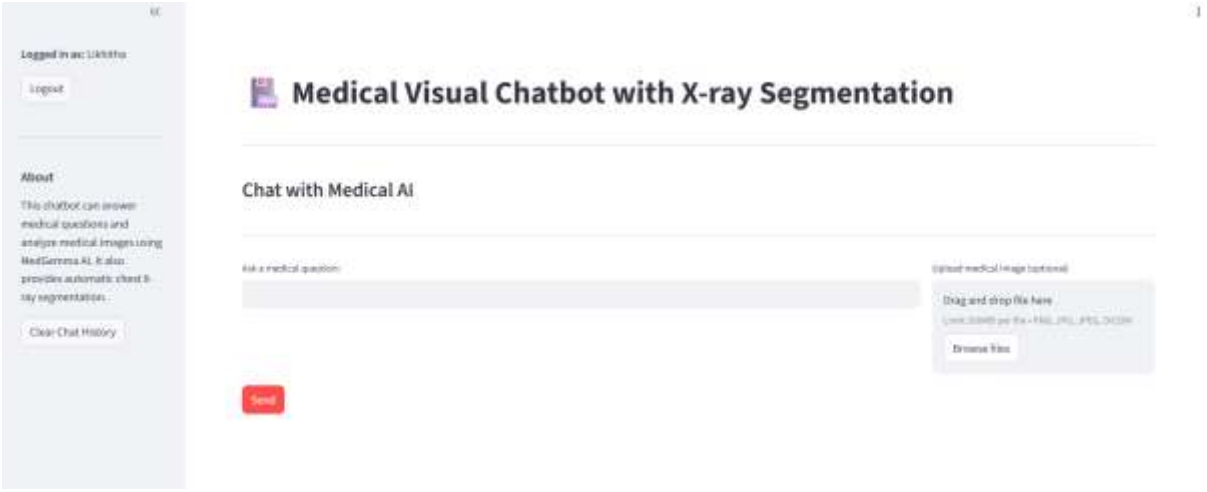
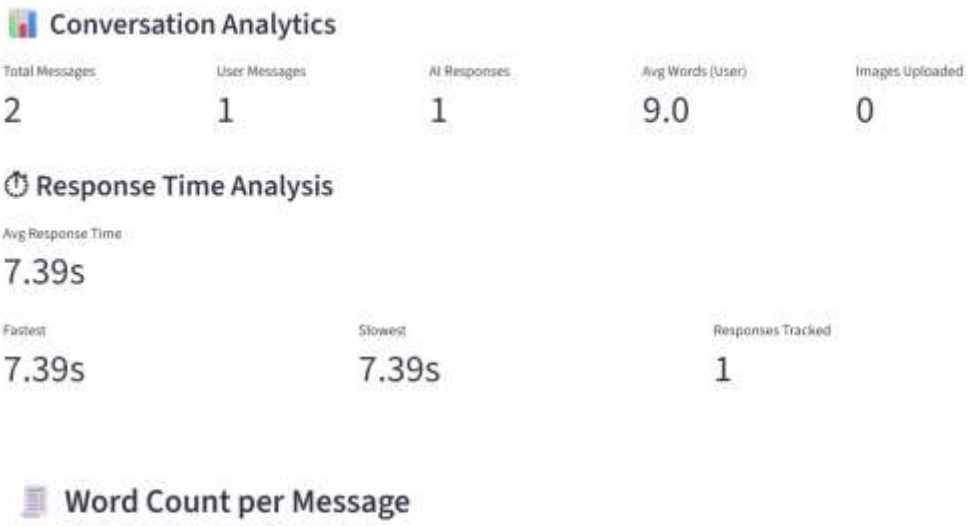


Fig 5.3: User Interface of the MediXpert Web Application

The MediXpert system is designed to process non-image-related medical inputs as well as image-related medical inputs, ensuring a more complete and interactive diagnostic framework. In non-image-related medical inputs, the system allows users to provide clinical details such as patient symptoms, medical history, age, or other relevant observations through its streamlined web interface. This information can either supplement the image analysis or be used independently to generate AI-based assessments and medical suggestions. By combining both visual and textual data, MediXpert enables a broader and context-aware diagnostic process, making it valuable in cases where imaging data is unavailable or when a deeper understanding of patient context is required. Its conversational and adaptive interface also allows users to pose general medical queries, receive AI-generated explanations, and explore related insights—positioning MediXpert as a versatile and educational healthcare assistant that supports both image and non-image-based analysis. The layout and functionality of the non-image-related input interface are illustrated in Fig 5.4.

Fig 5.4: Non-Image-Related Input Interface of the MediXpert System

The Conversation Analytics Dashboard as shown in Fig. 5.5 provides valuable insights into the chatbot's performance during non-image (text-only) interactions. It evaluates key metrics such as response time, word count per message, and user versus AI word contribution. The Response Time Analysis assesses the system's efficiency in generating replies, indicating how quickly the AI processes user queries. The Word Count per Message metric visualizes the length and complexity of both user inputs and AI responses throughout the conversation, reflecting engagement and message detail. Additionally, the user vs AI Word Contribution graph compares the total number of words exchanged between the user and the AI, helping to identify conversational balance and interaction dynamics. Together, these analytics provide a comprehensive understanding of communication patterns and the AI's responsiveness during text-only exchanges.



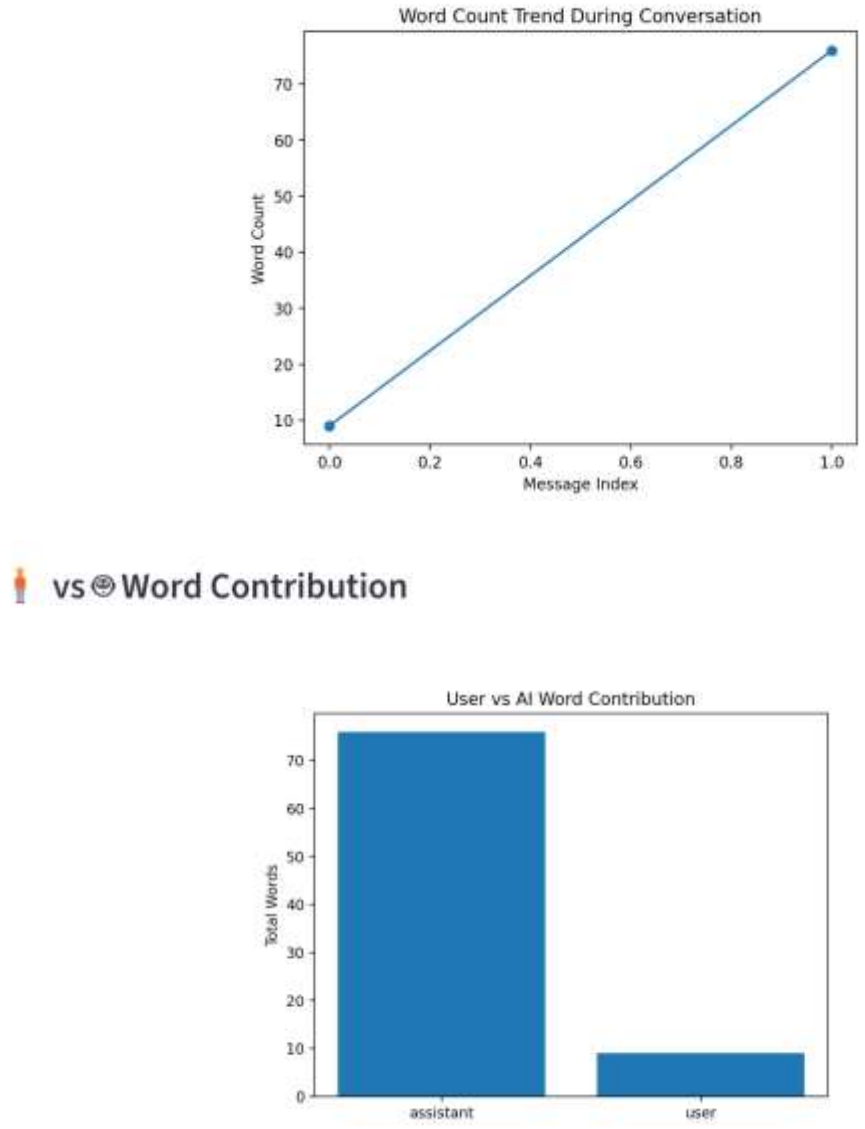


Fig 5.5: Conversation Analytics Dashboard — Non-Image Interactions

For image-related analysis, the MediXpert system focuses on processing and interpreting medical X-ray images using advanced deep learning models capable of identifying a wide range of thoracic abnormalities. Through the Streamlit-based interface, users can upload chest X-ray scans, which are then analyzed by the AI model to detect potential disease patterns and highlight areas of concern. Once an image is uploaded, the system automatically carries out preprocessing, feature extraction, and inference, generating comprehensive outputs such as predicted disease probabilities and visual explanations in the form of heatmaps. These outputs are especially useful for identifying and understanding conditions like pneumonia, pulmonary edema, fibrosis, pleural effusion, cardiomegaly, atelectasis, and infiltration, among other chest-related abnormalities. The heatmaps visually emphasize the specific regions of the lungs or thoracic cavity that contributed most to the

AI’s interpretation, making the diagnostic process more transparent and informative.

The interface ensures a seamless and intuitive workflow—from image upload to result visualization—allowing healthcare professionals, researchers, and students to interact directly with AI-driven diagnostics in real time. This capability makes MediXpert an effective and accessible platform for exploring how artificial intelligence can assist in the detection and interpretation of diseases in chest X-rays. The overall process and results of the image-related analysis are illustrated in Fig 5.6.

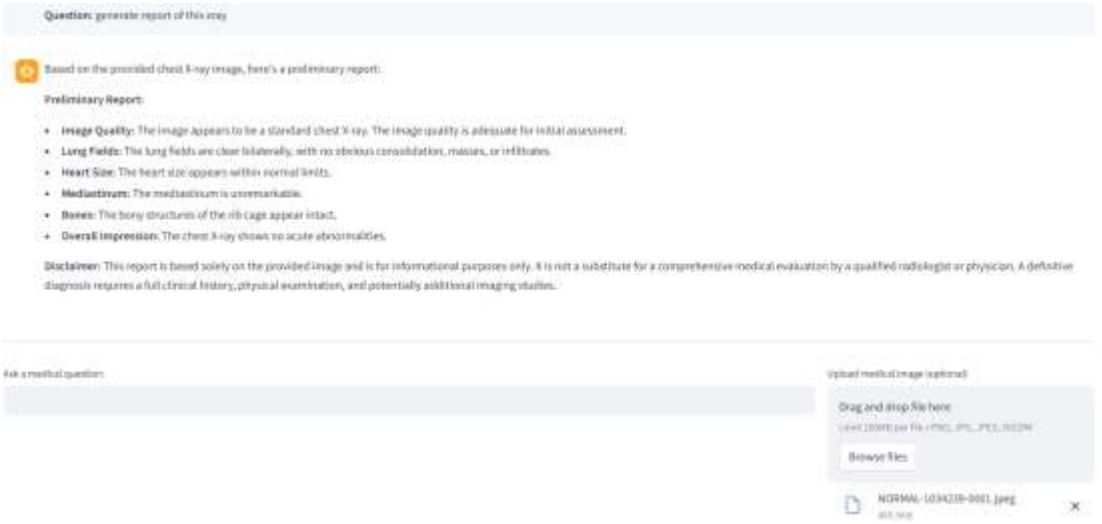


Fig 5.6: Image-Related Input Interface of the MediXpert System

The MediXpert system also provides a detailed disease classification output, which displays the likelihood or probability of various thoracic diseases detected in the analyzed X-ray image. After an image is uploaded and processed, the system generates a comprehensive list of possible conditions—each associated with a percentage score that represents the AI model’s confidence level. These classifications include a wide range of chest-related abnormalities such as atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, effusion, pneumonia, pleural thickening, cardiomegaly, nodule, mass, hernia, lung lesion, fracture, lung opacity, and enlarged cardiomediastinum.

Fig 5.7: Disease Classification Output of the MediXpert System

This structured presentation allows users to quickly interpret which conditions are most likely present, facilitating clinical review and comparison. It is emphasized that these probabilities serve as AI-generated indicators rather than definitive diagnoses. The disease classification results generated by the MediXpert system are illustrated in Fig 5.7.

The Conversation Analytics Dashboard as shown in Fig. 5.8 illustrates the chatbot's performance during image-based interactions. It analyzes metrics such as response time, word count per message, and user vs AI word contribution, along with image upload frequency. The dashboard evaluates how effectively the AI interprets both textual and visual data. It highlights the depth of responses and engagement during medical image analysis. Overall, it provides insights into the system's efficiency and multimodal understanding.

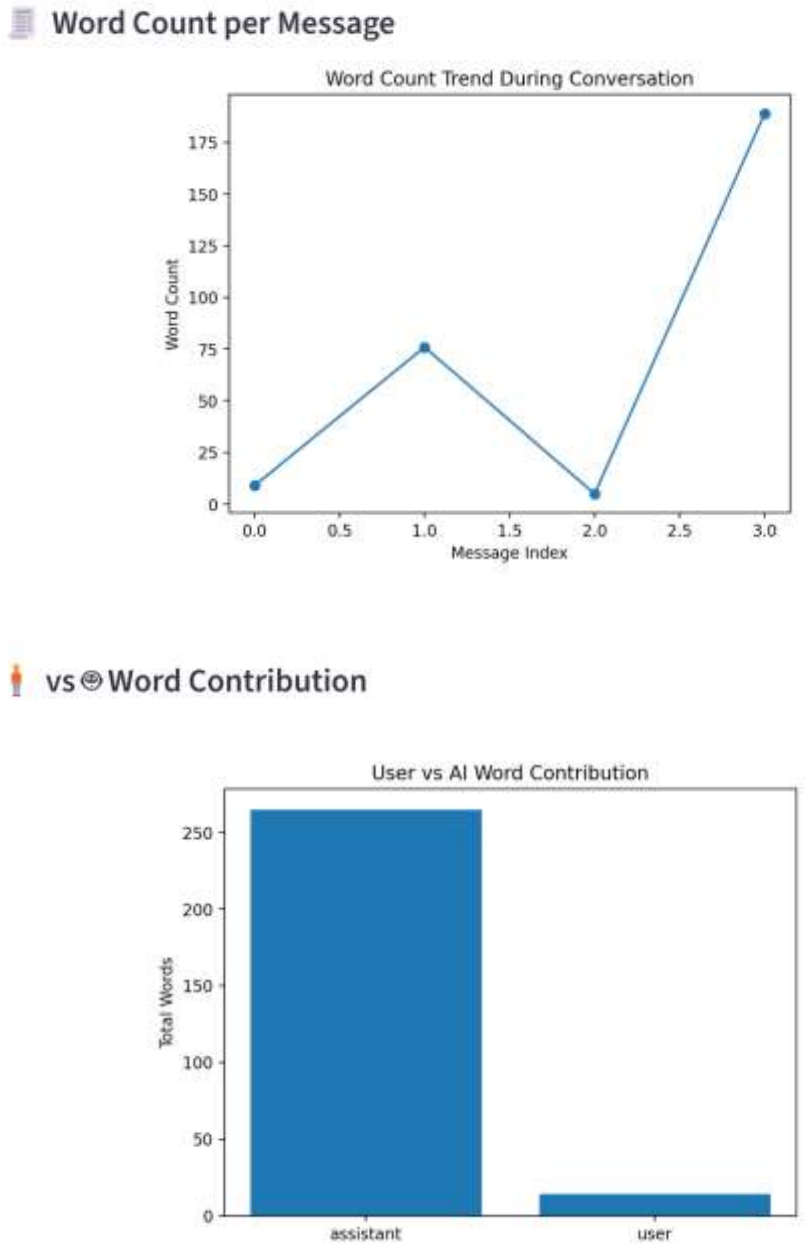


Fig 5.8: Conversation Analytics Dashboard — Image-Based Interactions

In addition to the above metrics, the dashboard also presents Image Upload Frequency and Detailed Message Statistics for the entire conversation as shown in Fig 5.9. These insights help visualize how often images were incorporated and how the dialogue evolved across multiple interactions. By analyzing the overall distribution of messages and image usage, the system provides a clear understanding of user engagement patterns. This comprehensive overview supports better evaluation of the chatbot's multimodal performance and interaction consistency throughout the session.

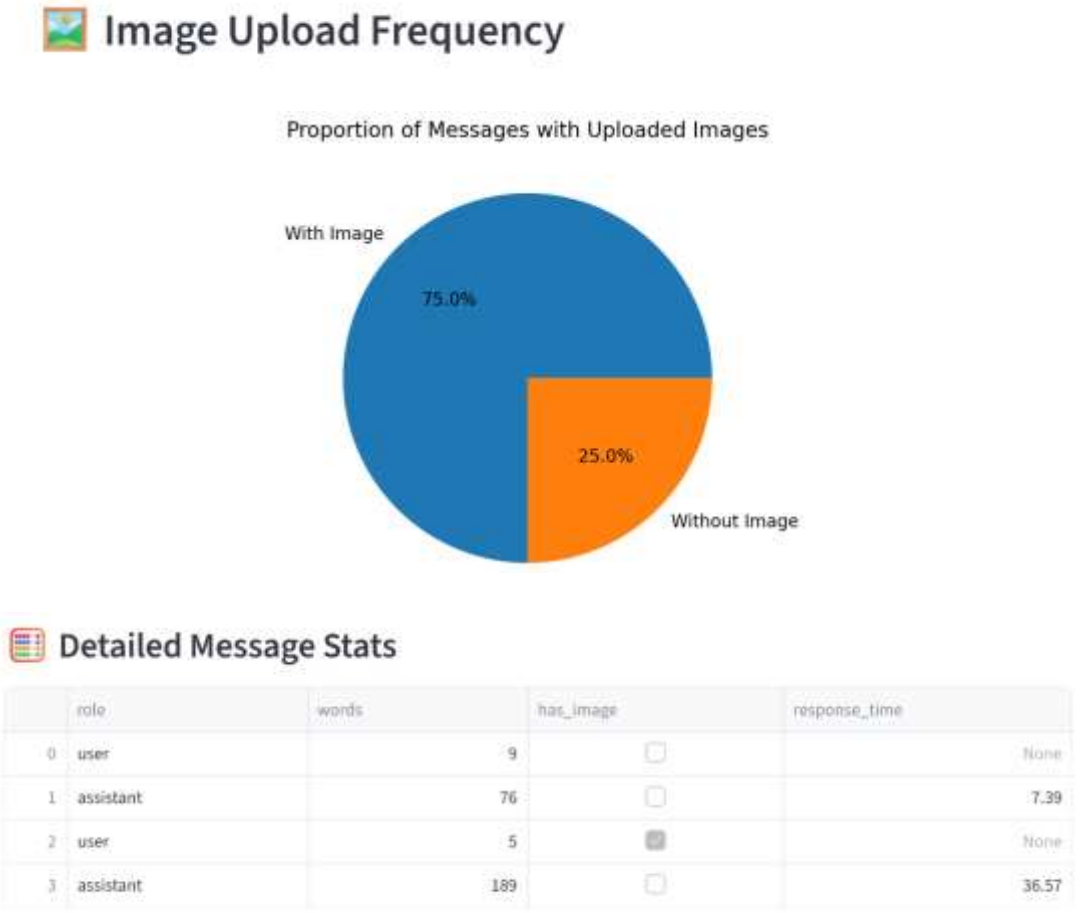


Fig 5.9: Overall Image Upload and Message Statistics

6 Conclusions and Future Enhancements

MediXpert successfully demonstrates the potential of multimodal artificial intelligence in clinical decision support. By integrating Natural Language Processing (NLP) for symptom understanding and Computer Vision (CV) for medical image interpretation, the system bridges the gap between text-based and image-based diagnosis. Through models such as Gemma, MedGemma, and Qwen2.5 for language comprehension, and ResNet, EfficientNet, and MedGemma for image analysis, MediXpert achieves a unified workflow that mirrors the holistic reasoning of healthcare professionals.

The implemented web-based interface enables seamless input of symptoms and image uploads, allowing users to receive diagnostic guidance in a conversational format. This innovation has significant implications for telemedicine, rural healthcare access, and early diagnosis, especially in environments with limited clinical resources. The project reinforces the Sustainable Development Goals (SDG 3 – Good Health and Well-Being; SDG 9 – Industry, Innovation, and Infrastructure) by promoting equitable access to intelligent healthcare technologies.

In the future, MediXpert can be enhanced through deeper integration with Electronic Health Records (EHRs) to incorporate patient history and lab data, thereby improving contextual accuracy in clinical recommendations. The performance of both language and vision models can be refined by fine-tuning them on larger, domain-specific datasets such as MIMIC-CXR and CheXNet, ensuring more precise and explainable outcomes. Incorporating Explainable AI (XAI) techniques would allow clinicians to visualize how predictions are derived, enhancing trust and transparency. Moreover, adding multilingual support will make the system more inclusive for non-English users, especially in rural and global telehealth environments.

The development of a mobile version could enable real-time diagnostics and triage in remote areas with limited access to computers. Lastly, integration with IoT-enabled diagnostic devices could allow continuous monitoring and automated alerts, transforming MediXpert into a comprehensive and intelligent clinical assistant that evolves alongside future advancements in healthcare technology.

References

1. Alibaba Cloud. (2024). Qwen2.5 Technical Report: Advanced Vision-Language Capabilities for Multimodal AI. arXiv preprint arXiv:2408.08258.

2. Che Liu Lee, J., Yoon, W., Kim, S., et al. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.

3. Douglas Townsell, R., Sun, L., Xia, Y., Qin, T. (2022). BiomedGPT: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. arXiv preprint arXiv:2305.17100.

4. Google Research. (2024). Gemma and MedGemma Models: Scalable Foundation Models for Multimodal Biomedical Tasks. Google DeepMind Technical Report, February 2024.

5. Ji Seung Ryu, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

6. Johnson, A. E. W., Pollard, T. J., Shen, L., et al. (2019). MIMIC-CXR: A large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042.

7. Pengyu Wang, X., Peng, Y., Lu, L., et al. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

8. Radford, A., Kim, J. W., Xu, T., et al. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning (ICML).

9. Rajpurkar, P., Irvin, J., Zhu, K., et al. (2017). CheXNet: Radiologist-Level Disease Detection on Chest X-Rays with Deep Learning. arXiv preprint arXiv:1711.05225.

10. Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971.

11. Xu, Y., Li, Z., Yang, Z., et al. (2024). Vision-Language Models in Medicine: A Review of Methods and Applications. Nature Digital Medicine, 7(3), 212–228.

12. Zhang, H., Zhang, P., Hu, X., et al. (2023). LLaVA: Large Language and Vision Assistant. arXiv preprint arXiv:2304.08485.

Glossary

- **Vision-Language Models (VLMs)** – AI models that jointly process visual and textual inputs to perform tasks such as image captioning, question answering, and medical diagnosis.
- **Transformer Models** – Neural network architectures (like GPT, BERT, and Gemma) designed for sequence-based tasks such as language understanding and translation.
- **Convolutional Neural Network (CNN)** – A deep learning architecture commonly used for analyzing and classifying image data.
- **Clinical Decision Support System (CDSS)** – A digital system designed to assist healthcare professionals in making evidence-based clinical decisions.
- **Explainable AI (XAI)** – An AI approach that makes model decisions transparent and understandable to human users.
- **ChromaDB** – A vector database used for storing and retrieving multimodal embeddings efficiently in real-time applications.
- **ResNet / EfficientNet** – Deep learning architectures designed for high-accuracy image classification and medical imaging tasks.
- **Electronic Health Records (EHRs)** – Digital versions of patients' medical histories, including symptoms, diagnoses, medications, and test results.
- **MIMIC-CXR / CheXNet Datasets** – Publicly available datasets of labeled chest X-rays used for training and evaluating medical image classification models.
- **Natural Language Processing (NLP)** – A subfield of AI that helps computers understand, interpret, and generate human language in a meaningful way.
- **Computer Vision (CV)** – A field of AI that enables machines to interpret and analyze visual information from images or videos.
- **Multimodal Integration** – The process of combining different types of data (like text, images, and audio) to create more context-aware and intelligent systems.
- **Fine-Tuning** – The process of adapting a pre-trained AI model to a specific domain (such as medical text or imaging) by training it on specialized datasets for improved accuracy and relevance.
- **Multimodal Embeddings** – Unified numerical representations that combine information from different data types (text, images, audio) into a shared feature space, allowing AI models to perform cross-modal understanding and reasoning.

Appendix

- **Architecture Design –** The system follows a modular design combining Natural Language Processing (NLP) and Computer Vision (CV) pipelines, unified through multimodal embedding fusion using transformer-based architectures and ChromaDB.
- **Workflow Pipeline –** Symptom descriptions and medical images are processed simultaneously: text is tokenized and encoded through language models, while images are analyzed using CNNs like ResNet and EfficientNet. Both embeddings are fused to generate diagnostic insights.
- **User Interface –** A Streamlit-based web interface allows users to enter symptoms, upload diagnostic images, and receive AI-generated clinical suggestions interactively.
- **Model Evaluation –** The system's performance was validated using sample datasets such as MIMIC-CXR and CheXNet, ensuring consistency between image interpretation and symptom analysis.
- **Deployment Environment –** MediXpert is deployed as a web application capable of local and cloud execution, with Python 3.x, PyTorch, and Streamlit forming the core environment stack.
- **Security and Ethics –** Patient data and medical images are anonymized before processing, ensuring compliance with data privacy and ethical healthcare AI standards.
- **Data Sources** – The project utilized publicly available medical datasets such as *MIMIC-CXR*, *CheXNet*, and *ChestX-ray8* for model training and evaluation of diagnostic accuracy.
- **Model Fine-Tuning Process** – Pre-trained models such as Gemma, MedGemma, and Qwen2.5 were fine-tuned on domain-specific datasets to improve clinical relevance and multimodal reasoning.
- **Limitations and Constraints** – The prototype currently relies on static datasets and lacks live clinical validation; further data diversity and real-world testing are required for production deployment.