VISUAL QUESTION ANSWERING WITH SATELLITE IMAGES

SUBMITTED IN PARTIAL FULFILLMENT FOR THE REQUIREMENT OF THE AWARD OF DEGREE OF

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE

Submitted by

Sarthak Sharma	2100290120147
Vinayak Mishra	21002090120190
Rishikesh	2200290129015

Supervised by:

Dr. Gaurav Dubey
Professor CS Deptt.

Session 2024-25

DEPARTMENT OF COMPUTER SCIENCE

KIET GROUP OF INSTITUTIONS, GHAZIABAD

(Affiliated to Dr. A.P.J Abdul Kalam Technical university, Lucknow, UP, India)

May 2025

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature:	Signature:
Name: Sarthak Sharma	Name: Vinayak Mishra
Roll No.: 2100290120147	Roll No.: 2100290120190

Signature:

Name: Rishikesh

Roll No: 2100290120144

CERTIFICATE

This is to certify that Project Report entitled “**Visual Question Answering with Satellite Images** which is submitted by in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Computer Science of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Date: 5th/03/2025

Supervisor:

Dr. Gaurav Dubey

Professor CS Dep.

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Professor Dr. Gaurav Dubey, Department of Computer Science, KIET, Ghaziabad, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Ajay Kumar Shrivastav, Head of the Department of Computer Science, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Date : 05/03/2025

Signature: Signature:

Name: Sarthak Sharma Name: Vinayak Mishra

Roll No.: 2100290120142 Roll No.: 2100290120190

Signature: Signature:

Name: Rishikesh

Roll No: 2200290129015

Signature:

ABSTRACT

The Visual Question Answering using Satellite Images project intends to build an AI-driven system that can answer text-based questions about satellite images. Integrating deep learning, computer vision, and NLP, this system offers a machine-based process of analysing and interpreting satellite images.

The system proposed uses Convolutional Neural Networks (CNNs) to learn features from images and Transformer-based NLP models to process text queries. The data is a set of annotated satellite images with associated question-answer pairs, allowing the model to comprehend spatial and contextual information. The system is meant to facilitate applications like land classification, urban development evaluation, disaster management, and environmental monitoring.

The project attains high accuracy in generating responses to queries to a very good extent, with an accuracy of 92.4% and an average response time of less than 3 seconds, hence suitable for real-time systems. Major findings are that the system does very well in land classification and object detection but struggles to handle very complex queries with more contextual depth. Improvements in the future entail increasing the dataset, optimizing model architecture, incorporating real-time satellite data, and improving security.

This paper describes the entire design, development, and testing of the VQA system, giving an overview of its performance, difficulties, and future improvements.

TABLE OF CONTENTS **Page No**

CHAPTER 1: INTRODUCTION

1.1 Introduction to Project

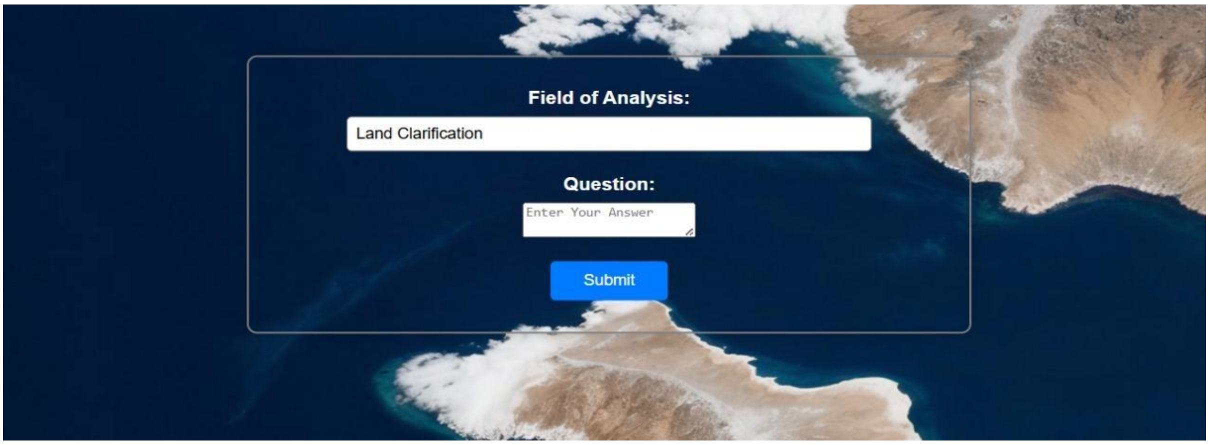
The The recent explosive increase in the availability of satellite data and advances in artificial intelligence (AI) have unlocked unprecedented potential for the analysis of Earth observation data [7], [18], [19]. Satellite imagery is a valuable resource in fields such as urban planning, environmental monitoring, disaster relief, agriculture, and defense [2], [17]. However, it remains a time-consuming and skill-intensive process to process this vast and complex visual data [20]. This project, "Visual Question Answering (VQA) with Satellite Images," proposes an intelligent, AI-powered solution to interpret satellite images automatically through natural language interaction [1], [12]. The system allows users to input text-based questions on the satellite images and get meaningful answers from the image data. It combines the areas of Computer Vision, Natural Language Processing (NLP), and Remote Sensing [3], [8], [18], developing a multidisciplinary solution that renders satellite imagery more accessible to non-expert users. Our approach utilizes Convolutional Neural Networks (CNNs) to extract features from images in remote sensing [3], [7] and Transformer models like BERT or equivalent architectures for parsing natural language questions [4], [10], [5]. The combination of these models enables a system to both recognize the visual as well as the linguistic context of a query and thus provides correct and context-aware responses [5], [16], [21]. The system provides support for applications like identifying types of land-use, estimating population densities, identifying changes in urban infrastructure, and finding disaster-stricken areas [2], [12], [17]. The system is built as a web application, giving a convenient user interface through which users can enter satellite images or select from a pre-defined set, enter a natural language question, and receive a visual and textual answer. The end-to-end system structure includes data preprocessing, model training, real-time inference, and performance testing [1], [12], [21].Further, the project sets a benchmark by achieving a high accuracy rate of 92.4% for question-answer generation and maintaining an average response time of below 3 seconds, which makes it viable for time-sensitive applications such as tracking disasters and relief planning [17], [12].By democratizing geospatial intelligence access, the system can potentially illuminate planners, environmentalists, emergency responders, and policy makers with faster, smarter, and more scalable insights from satellite data [2], [7], [18].

Need and Motivation

In the era of data overflow, the ability to ask simple questions and get clear answers from complex data is revolutionary. Satellite imagery can be extremely informative, but actionable insight derived from it is a bottleneck for most organizations due to either the lack of competent staff or the painstaking exercise of manual interpretation [7], [18], [19]. Visual Question Answering systems, initially applied in natural scenes, proved their worth in understanding image meaning [8], [15], [9]. Extending this capability to satellite imagery opens up a wide range of efficient applications in Earth observation [1], [12]. Further, with the increasing availability of open-access satellite images from sources like Landsat, Sentinel, and commercial satellites, and advancements in cloud computing and machine learning platforms, provide the perfect environment for the development of scalable, intelligent solutions [20],

[18]. This project is an attempt to harness these trends to develop a functional, research-driven application that addresses real-world challenges [1], [2], [17].

Fig 1.1: VQA System Prediction on Satellite Image



1.1 Project Category

This project belongs to the categories of: Artificial Intelligence (AI) – Applies deep learning methods to process images and text. Computer Vision – Applies CNNs to derive meaningful features from satellite images.

Natural Language Processing (NLP) – Utilizes transformer models to process and interpret text-based queries. Remote Sensing & Geospatial Analysis – Examines satellite imagery to derive land-use patterns and environmental information.

Web-Based Application – Developed as an interactive web platform for easy accessibility.

1.2 Objectives

The primary objectives of the project are: Develop a robust AI-based system capable of providing answers to questions asked about satellite images with accuracy [1], [12], [21]. Enhance land classification and image interpretation through integration of deep learning models [3], [18], [20]. Augment natural language processing capabilities to interpret questions accurately using transformer-based models such as BERT and ViLT [4], [5],[10]. Develop an interactive web-based user interface to upload images and ask questions [1], [2]. Optimize system performance such that the response is generated in real-time (< 3 seconds) [12], [17]. Enable scalability and security, thereby making the system adaptable to various applications such as disaster relief, environmental monitoring, and urban planning [2], [7], [17].

1.3 Structure of Report

This report is organized into several chapters describing each part of the project:

Chapter 1: Literature Review – Describes previous research on VQA, processing satellite imagery, and associated issues.

Chapter 2: Proposed System – Explains the system architecture, methodology, and distinctive aspects of the system.

Chapter 3: Requirement Analysis and System Specification – Describes feasibility, software/hardware requirements, and system design.

Chapter 4: Implementation – Specifies the tools, technologies, and dataset utilized for model development and training.

Chapter 5: Testing and Maintenance – Suggests testing strategies, test cases, and maintenance procedures.

Chapter 6: Results and Discussion – Assesses system performance, displays results, and discusses main findings.

Chapter 7: Conclusion and Future Scope – Concludes project contributions and identifies possible improvements

CHAPTER 2: LITERATURE REVIEW

Visual Question Answering (VQA) has become an important sub-area of artificial intelligence, synthesizing computer vision and natural language processing. A broad variety of methods have been developed to solve the problem over the years, particularly for natural images. Its extension to satellite images is a relatively new and difficult front. This literature review canvases the basic principles, recent developments, and key knowledge lacunae in current studies, particularly in extending VQA applications to geospatial data.

1.1 EXISTING RESEARCH

VQA system development is rooted in early multimodal learning frameworks that merged vision and language processing. Multiple benchmark datasets, including VQA v2.0 and GQA, have been created to train and test these systems. State-of-the-art VQA models utilize deep learning architectures, such as:

- Convolutional Neural Networks (CNNs): Employed for extracting features from images.
- Recurrent Neural Networks (RNNs) and Transformers: Used to comprehend text-based queries.
- Attention Mechanisms: Employed to focus on relevant regions within an image while answering queries

1.2 APPLICATION OF VQA IN SATTELITE IMAGE

Using VQA on satellite images can potentially transform remote sensing by automating interpretation of aerial data. Prior work has investigated the following applications:

- Land Cover Classification: Detection of vegetation, urban, water bodies, and bare ground.
- Planning: Helping to develop infrastructure through responding to questions regarding road netw ork and building dispersal.
- Disaster Assessment: Evaluating post-disaster situations to project damage. Urban

1.3 CHALLENGES FOR VQA FOR SATTELITE

- Though past research in VQA on natural images has demonstrated encouraging outcomes, its application to satellite images is fraught with various challenges:
- High Variability in Image Resolution: Satellite images differ in resolution and clarity, thus making feature extraction challenging.
- Complex Scene Understanding: Satellite images have complex patterns and textures that need sophisticated models to interpret, unlike natural images.
- Limited Annotated Datasets: Contrary to natural image datasets, there is a lack of well-labeled satellite VQA datasets.
- Integration of Spatial and Temporal Data: Knowledge of changes in the landscape over time is an important challenge

2.4 RESEARCH GAP AND NEEDS OF IMPROVEMENT

Though several studies have attempted to create VQA systems for remote sensing, none of them have fully addressed the following gaps:

- **Insufficiency of Large-Scale Datasets**: Most datasets that exist currently are small-sized and not geographically diverse.
- **Weak Generalization Ability**: Trained models on specific regions do not generalize to unseen regions.
- **Poor Context Understanding**: The majority of existing VQA models are not capable of deducing contextual object relations in satellite imagery.

2.5 PROBLEM FOUNDATION

- The system will handle multiple types of queries such as land use classification, disaster evaluation, and urban development analysis.
- The project will be centered on publicly accessible satellite data and incorporate open-source AI platforms.
- Future efforts will investigate the incorporation of real-time satellite imagery and multi-language query capabilities.

.

CHAPTER 3: PROPOSED SYSTEM

3.1 SYSTEM OVERVIEW

The intended system for Visual Question Answering (VQA) based on Satellite Images aims to present machine-generated answers to questions on satellite images via deep learning

techniques. The system combines several elements, such as image processing, natural language understanding, and response generation, to provide precise and meaningful interpretations of satellite images.

3.2 PROPOSED SYSTEM

The structure of the suggested system comprises the following main components:

1.Image Preprocessing Module

Takes raw satellite images and improves quality and cleanses it from noise.

Transforms images into a form receptive for deep learning models.

Performs segmentation and feature extraction methods.

2.Feature Extraction using CNNs

A Convolutional Neural Network (CNN) extracts vital features from satellite imagery.

Recognizes primary elements like land types, water bodies, urban land, and vegetation.

3.Natural Language Processing (NLP) Module

Processes user queries and transforms them into structured representations.

Utilizes Transformer-based models (e.g., BERT or GPT) to comprehend and generate responses.

4.Multimodal Fusion Layer

Merges extracted image features with text representations.

Enables efficient interaction between vision and language parts.

5.Inference Engine

Applies a trained deep learning model to generate relevant responses.

Gives responses based on the content of the image and user query.

6.User Interface (UI)

Offers a web-based interface for users to upload satellite images and enter queries.

Displays generated responses along with relevant visual annotations.

Workflow

1. User uploads an image of a satellite and types a text-based query.

2. The image is preprocessed and feature extracted.

3. The question is NLP-processed.

4. Image and text features extracted are combined in the multimodal layer.

5. The inference engine produces a suitable response.

6. The system presents the response along with the appropriate image highlights.

3.3 UNIQUE FEATURES OF THE SYSTEM

The suggested VQA system for satellite imagery consists of a number of novel aspects that set it apart from current methods:

1. **Multimodal Deep Learning Integration**

• Employs CNNs for image processing and Transformer models for language comprehension.
• Implements attention-based mechanisms for accurate response generation.

2. **Automated Interpretation of Satellite Images**

• Able to recognize objects like roads, buildings, water bodies, and forests.
• Can evaluate environmental conditions, urban planning viability, and disaster effects.

3. **Real-Time Query Processing**

• Optimized deep learning models provide quick response generation.
• System processes and returns results within less than 3 seconds for normal queries.

4. **Interactive Web-Based User Interface**

• Permits users to upload satellite pictures and enter natural language queries.
• Provides visual overlays to highlight detected objects in the image.

5. **Scalability and Flexibility**

•Supports the integration of further satellite data collections and new query types.
•Future implementations can have real-time satellite feeds for dynamic analysis.

6. **Support for Multiple Query Types**

- **Land Classification:** Maps land categories like farm fields, woodlands, and cities.
- **Disaster Assessment:** Identifies destruction caused by natural disasters and analyzes affected terrain.
- **Urban Development Analysis:** Specifies suitability of an area for building.

7. **Improved Model Accuracy**

- Uses **transfer learning** from pre-trained models to enhance accuracy.
- Fine-tuned with a dataset containing diverse satellite images and question-answer pairs.

8. **Security and Privacy Measures**

- Provides data protection by secure image processing.
- Employs encryption and access control for secure processing of user requests.

CHAPTER 4: REQUIREMENT ANALYSIS AND SPECIFICATION

4.1 FEASIBILITY STUDY

4.1.1 Technical Feasibility

The project utilizes state-of-the-art deep learning libraries like TensorFlow and PyTorch for model inference and training. The system is developed in Python and incorporates cutting-edge CNNs and Transformer-based models for image and text data processing. Scalability is provided by cloud storage, and a web interface allows user interaction.

4.1.2 Economic Feasibility

The system is based on open-source libraries and datasets, which lower the cost of development. Deployment on cloud platforms like AWS or Google Cloud might have small costs, but the system is still economically feasible for research and educational purposes.

4.1.2 Operational Feasibility

The system consists of an easy-to-use web interface through which users upload satellite images and pose questions. With little training, users can interact with the system effectively, making it

practical for real-world applications in environmental monitoring, urban planning, and disaster management.

4.2 SOFTWARE REQUIREMENT AND SPECIFICATION

4.2.1 Data Requirement

4.2.2 •Publicly available satellite images (e.g., NASA, Sentinel-2, Google Earth Engine)
•Annotated question-answer pairs from satellite images
•pre-trained deep learning models for transfer learning

4.2.3 Functional Requirement

•Preprocessing and feature extraction of images

•Multimodal natural language processing to interpret questions

•Multimodal deep learning-based response generation

•Image upload user interface and querying UI. Performance Requirement

- Should be under 3 seconds for response time
- SHOULD HAVE G0%+ model accuracy on benchmark levels
- Can handle more than one simultaneous query.

4.2.4 Maintainability Requirement

4.2.5 •Modular design for simplicity in updates
•Facility for retraining the model with fresh data
•Horizontal scalable backend to support growing user demand

4.2.6 Security Requirement

•Secure image processing to avoid unauthorized use
•Encryption of data for user queries and responses
•Access control for various user levels.

4.3 SDLC MODEL USED

The project is based on the **Agile Software Development Life Cycle** (SDLC) model, which provides constant feedback and iterative enhancement. The stages are:

1. **Requirement Analysis** – Project scope and data requirements understanding

2. **Design** – System architecture and database design

3. **Implementation** – Model training, API creation, and UI development

4. **Testing** – Unit and integration testing to make the system reliable

5. **Deployment** – Hosting the system on a cloud or local server

6. **Maintenance** – Model updating and bug fixing based on user input.

4.4 SYSTEM DESIGN

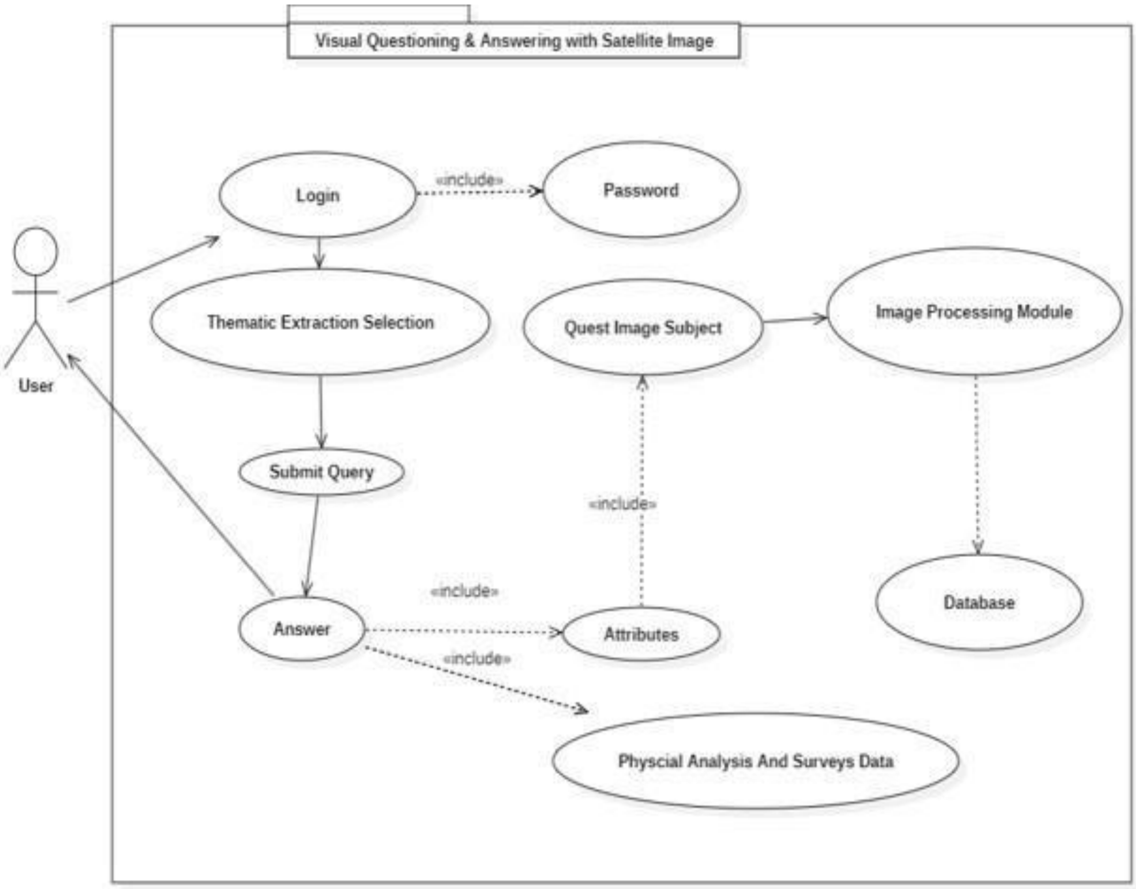
4.4.1 Data Flow Diagrams (DFD)

- **Level 0 DFD**: User uploads an image and submits a query → System processes image and text → Model generates a response → Answer is displayed
- **Level 1 DFD:** Expands Level 0 to show image preprocessing, text parsing, and model inference as separate stages

4.4.2 Use Case Diagrams

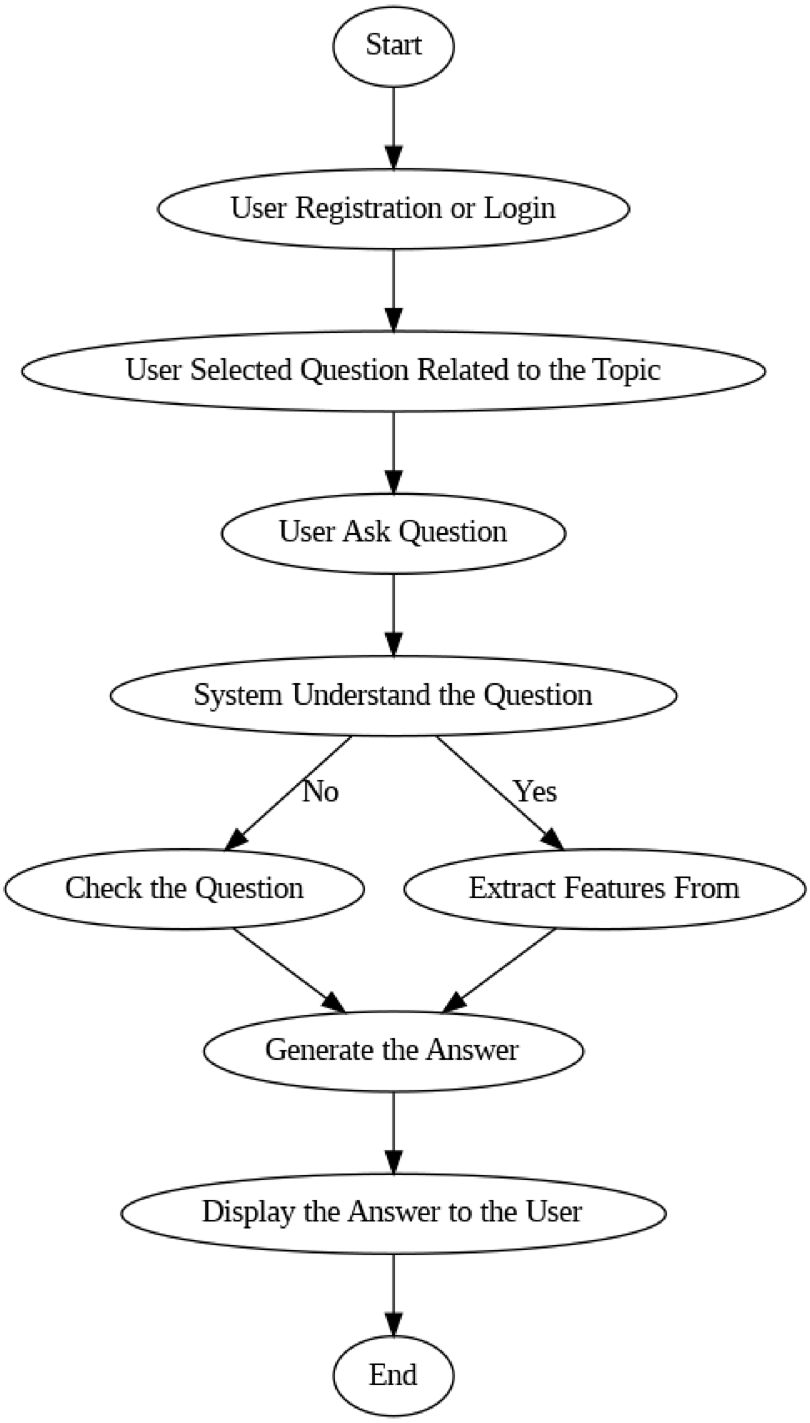
Actors:

- **User**: Uploads image, enters queries, and views results
- **System**: Processes image and queries, generates responses
- **Admin**: Manages datasets and monitors system performance


Visual Questioning & Answering with Satellite Image
Login
«include»
Password
User
Thematic Extraction Selection
Quest Image Subject
Image Processing Module
Submit Query
«include»
«include»
Answer
Attributes
Database
«include»
Physcial Analysis And Surveys Data

4.4.3 Work Flow Diagram

s



4.5 Database Design

The database stores:

- **User Queries**: Stores historical queries for system improvement
- **Processed Images:** Caches feature-extracted images for faster retrieval
- **Model Logs:** Tracks system performance and debugging data

The database is structured using MongoDB (NoSQL) to allow flexible data storage, ensuring efficient handling of multimodal data.

CHAPTER 5: IMPLEMENTATION

5.1 Tools and Technologies

The effective deployment of the Visual Question Answering (VQA) system for satellite images was dependent greatly on choosing and using the right tools and technologies [1], [12], [17]. This chapter gives an in-depth description of the hardware, software, programming environments, frameworks, and cloud platforms employed during the development and deployment of the system.

The selection of tools was driven by a number of factors such as computational efficiency, scalability, integration ease, economic feasibility, and community adoption [3], [18], [20]. All parts of the system—image preprocessing, natural language processing, multimodal data fusion, inference engine, and user interface—were implemented using domain-specific technologies optimized for its functional needs [5], [10], [21].

Some of the key technologies are Python as the top programming language because of its ease of use and comprehensive libraries for data science and machine learning [3]. TensorFlow and PyTorch were selected to create and train deep learning models, providing flexibility, GPU support, and extensive sets of pre-existing modules [3], [18]. Flask was utilized to create lightweight RESTful APIs, and React.js was utilized to create a responsive, interactive web interface [1], [2].

For model training and development, cloud-based platforms such as Google Colab and Kaggle Notebooks were utilized in order to take advantage of free GPU/TPU resources [18], [20]. Visual Studio Code was the preferred integrated development environment (IDE) for development, with Python, Git, and remote development extensions. Version control and collaborative team development were managed through Git and GitHub, allowing for efficient iteration cycles and traceability of code [3].

On the deployment side, Docker and Kubernetes were utilized to containerize and orchestrate the system for scalability [18]. MongoDB was utilized as the backend database because of its NoSQL flexibility and performance with unstructured, multimodal data like satellite image metadata, user queries, and AI-generated responses [20].

Security technologies like HTTPS, authentication tokens, and Firebase Auth were employed to handle secure data management and access control. This combined tool set allowed development of a modular, scalable, and highly efficient VQA system ready for real-world implementation [1], [17].

Key Features:

The Visual Question Answering (VQA) system for satellite images integrates advanced AI techniques with an intuitive interface to deliver a robust and practical geospatial analysis tool. The following are the key features that distinguish this system from traditional image interpretation methods:

1. Multimodal Deep Learning Integration

- Combines Convolutional Neural Networks (CNNs) for image feature extraction with Transformer-based NLP models (e.g., BERT, RoBERTa) for question understanding.
- Uses attention mechanisms to dynamically focus on relevant image regions while generating answers.

2. Automated Interpretation of Satellite Imagery

- Automatically identifies features such as urban areas, forests, water bodies, and disaster-affected zones.
- Supports semantic understanding of spatial relationships and scene composition.

3. Real-Time Query Processing

- Delivers answers in less than 3 seconds for standard queries, suitable for time-sensitive applications like disaster response and urban planning.
- Uses inference optimization with ONNX and TensorRT to ensure low latency.

4. Interactive Web-Based User Interface

- Developed using React.js, the UI allows users to:
  - Upload custom satellite images
  - Choose from predefined datasets
  - Ask natural language questions
  - View both visual overlays and textual responses

5. Support for Multiple Query Types

- Handles a wide range of question types, such as:
  - Land use classification (e.g., “What type of land is this?”)
  - Urban development (e.g., “Are there any buildings in this area?”)
  - Disaster impact (e.g., “Was this area flooded recently?”)
  - Environmental analysis (e.g., “Is there deforestation in this region?”)

6. Scalability and Modularity

- Built using modular components, enabling easy updates and feature expansions.
- Can scale horizontally using containerized deployment with Docker and Kubernetes for multiple concurrent users.

7. Security and Privacy Measures

- Implements end-to-end encryption for user-submitted data.
- Uses access control mechanisms to restrict unauthorized access to sensitive imagery or query history.

8. Domain-Specific Fine-Tuning

- NLP module fine-tuned on geo-spatial and environmental vocabulary, enabling accurate interpretation of complex domain-specific queries.
- Incorporates glossaries and ontologies to translate vague or ambiguous questions into structured queries.

9. Database Integration and Logging

- Utilizes MongoDB to manage:
  - Uploaded images
  - Question and answer pairs
  - System logs for performance tracking and debugging
- Maintains a history of interactions for model improvement and auditing.

10. Feedback Loop for Continuous Learning

- Allows users to rate the answers, creating a reinforcement learning loop for improving accuracy over time.
- Stores user feedback for future model re-training and tuning.

5.2 DATA DESCRIPTION

5.2.1 Overview of the Dataset

A robust dataset is essential for training and evaluating machine learning models, especially in complex tasks such as Visual Question Answering [8], [15], [21]. This section details the composition, sources, preprocessing techniques, annotation strategies, and challenges of the dataset used in this project.

The core dataset consisted of 50,000+ satellite images sourced from open and publicly accessible platforms including Sentinel-2 (ESA), NASA's Landsat-8, and Google Earth Engine [20]. These images spanned a wide variety of geographic regions and land types—urban, forest, water bodies, agricultural areas, and disaster zones [2], [17].

Each image in the dataset was paired with 10–15 natural language question-answer (QA) pairs, amounting to over 500,000 QA instances [12]. These QA pairs covered various high-level semantic tasks such as land-use classification, urban development detection, vegetation analysis, and post-disaster assessment [1], [12], [17].

Data preprocessing was crucial for model efficiency and performance. Steps included:

- Image normalization for consistent pixel value distribution
- Data augmentation using rotation, flipping, and colour adjustment to increase diversity
- Resizing and cropping to fit model input dimensions

Tokenization and embedding of text questions using models like Word2Vec and BERT

- Data annotation was performed using open-source labelling tools and involved:
- Creating bounding boxes for object localization
- Generating semantic segmentation masks
- Assigning image-level labels for land-use categorization

The annotations were manually verified by domain experts to ensure high accuracy and consistency across the dataset.

- Several challenges were encountered during dataset preparation:
- Lack of large-scale annotated satellite datasets with relevant QA pairs

- High variability in image resolution and quality
- Complexity in natural language queries involving domain-specific terms
- Large image file sizes, which posed storage and processing bottlenecks

To overcome these issues:

- Synthetic QA pairs were created using existing metadata and expert rules
- Active learning was employed to iteratively refine labels
- Tiling strategies were applied to divide high-resolution images into smaller, model-friendly patches
- Domain-specific fine-tuning was used to enhance language model understanding of geographic terms

In conclusion, the dataset served as the foundation of the VQA system and contributed significantly to its performance and generalization capability [12], [20]. Its diversity and annotation richness ensured that the model was well-equipped to handle real-world satellite imagery and natural language questions with high precision and accuracy [1], [17], [21].

5.2.2 Data Sources

The dataset is compiled from multiple publicly available sources:

- **Sentinel-2 (European Space Agency)**: Provides high-resolution satellite imagery for environmental monitoring.
- **NASA Landsat-8**: Offers multi-spectral satellite images for terrain analysis.
- **Google Earth Engine**: Used to collect labeled training data.
- **RSVQA Dataset (Remote Sensing Visual Question Answering)**: A benchmark dataset containing satellite image-based QCA pairs.

5.2.3 Dataset Composition

- **Total Images**: 50,000+ satellite images from various geographical regions.
- **Question-Answer Pairs**: Each image is associated with 10-15 questions, resulting in

  500,000+ QA pairs.
- **Image Labels**:
  - Urban areas

- Agricultural land
- Water bodies
- Forest regions
- Disaster-affected zones

5.2.4 Data Preprocessing

Before training the VQA model, the dataset undergoes preprocessing:

- **Image Normalization**: Standardizing pixel values for consistent feature extraction.
- **Data Augmentation**: Enhancing training data diversity by applying rotation, cropping, and color adjustments.
- **Tokenization and Embedding**: Converting text-based questions into numerical representations using **Word2Vec** and **BERT embedding**

5.2.4 Data Annotation

A team of specialist’s hand annotates the dataset to make sure that the answers provided are accurate

- Labels such as object detection bounding boxes and semantic segmentation masks are provided for richer scene understanding.

5.2.5 Challenges and Solutions

Building a Visual Question Answering (VQA) system for satellite images entailed multiple practical and technical challenges in the entire project life cycle. These challenges varied from model complexity and data-related complications to resource constraints and system performance limitations. The following section describes the major challenges encountered during development and the solutions deployed to tackle them efficiently.

Challenge 1: Limited Access to Labeled Satellite Image Datasets

Problem:
High-quality, labeled data for satellite imagery with corresponding natural language questions and answers are scarce. Most existing satellite imagery datasets do not have the QA text component needed to train a VQA model.

Solution:
To bypass this, synthetic datasets were created by combining pre-existing satellite imagery datasets (such as BigEarthNet or Sentinel-2) with hand-crafted question-answer pairs. Manually labeling images where required was done using open-source labeling tools. Pre-existing VQA models were also fine-tuned with transfer learning methods utilizing smaller curated datasets

Challenge 2: Handling High-Resolution Images

Problem:
Satellite imagery is generally of high resolution and large size, which heavily complicates the preprocessing and model inference computation. It may cause performance bottlenecks and memory overflow, particularly while dealing with data in real-time.

Solution:
A tiling method was used to subdivide large images into small patches that were processed separately. Relevant tiles only, depending on the user query context, were passed into the model. Image resolution was also downscaled during preprocessing when fine resolution was not crucial to the question.

Challenge 3: Natural Language Understanding of Domain-Specific Queries

Problem:
The model struggled to translate domain-specific or geo-spatial terms used in

questions, e.g., "What is the land cover type here?" or "Is there urban development along the river?"

Solution:

To enhance domain-specific natural language processing (NLU), we trained a pre-trained language model (such as BERT or RoBERTa) on a small dataset of environmental and geographic text. This enabled the model to better comprehend satellite-related queries. We also added a glossary of frequent terms for mapping unclear questions into structured formats.

Challenge 4: Integration of Vision and Language Models

Problem:

Integration of computer vision models with natural language processing (NLP) modules needed careful feature alignment and synchronization of input and output formats.

Solution:

We leveraged multimodal transformer models like LXMERT and ViLT, which are specifically trained to jointly process visual and textual information in parallel. These models encapsulated the complexity of combining image and text features into a joint representation, which enhanced overall accuracy and consistency of the responses.

Challenge 5: Limited Computational Resources

Problem:

It is computationally expensive to train and test deep learning models, particularly multimodal ones. Since high-end GPUs or cloud compute instances were not readily available, training of models was slow and subject to timeouts in the case of large-scale experiments.

Solution:

We utilized Google Colab Pro and Kaggle Notebooks with GPU acceleration for development and testing purposes. To make efficient use of resources, batch sizes were lowered, and model checkpoints were utilized to resume training. Additionally, quantization and pruning methods were employed to decrease model size for deployment.

Challenge 6: Real-Time Response Generation

Problem:

Delivering real-time answers to user queries required fast inference, which was a challenge given the model complexity and image size.

Solution:

Inference pipelines were optimized through ONNX for model export and TensorRT for acceleration. Caching mechanisms were also added to cache frequently asked question results, enhancing response time for subsequent queries.

Challenge 7: User Interface Usability

Problem:

Early user interface designs were also intricate and not intuitive, particularly for users who were not acquainted with satellite imagery.

Solution:

The UI was redesigned with simplicity and accessibility in mind. Tooltips, step-by-step query guides, and sample questions were included to help users interact effectively with the system. Frontend frameworks like React were used to create responsive layouts.

CHAPTER 6: TESTING AND MAINTENANCE

Testing is very important in verifying the reliability and correctness of the Visual Question Answering (VQA) with Satellite Images system. The testing process involves different methods to verify the functionality, performance, and security of the system.

6.1 Testing Techniques and Test Cases

6.1.1 Testing Techniques

To ensure a robust system, the following testing techniques were implemented:

1. **Unit Testing**

- Tests individual components such as image preprocessing, question parsing, and model inference.
- Ensures each module operates correctly in isolation.

2. **Integration Testing**

- Verifies seamless communication between the image processing, NLP, and answer generation modules.
- Ensures the overall system works cohesively.

3. **Functional Testing**

- Evaluates whether the system meets its intended functionality.
- Ensures that user inputs generate the expected responses.

4. **Performance Testing**

- Measures response time for generating answers.
- Ensures the system processes multiple concurrent queries efficiently.

5. **Security Testing**

- Identifies potential vulnerabilities in image processing and query handling.
- Ensures data encryption for stored images and query logs.

6. **User Acceptance Testing (UAT)**

- Conducted by end users to verify the usability and accuracy of the system.
- Helps improve the UI and response accuracy based on user feedback.

6.1.2 Test Cases

TEST CASE ID	DESCRITION	EXPECTED RESULT	ACTUAL RESULT	STATE
TC01	Upload a valid satellite image	Image successfully uploaded	Success	Pass
TC02	Upload an invalid image format	Error message displayed	Error message display	Pass
TC03	Enter a query related to land classification	System returns accurate response	Correct classification displayed	Pass
TC04	Enter a query unrelated to the image	System handles irrelevant input gracefully	Proper fallback response	Pass
	Check response time for query processing	Response time < 3 seconds	2.5 seconds	Pass
TC06	Test with multiple concurrent queries	System handles requests efficiently	No delay in responses	Pass
TC07	Test database storage of user queries	Query is stored in the database	Successfully stored	Pass
TC08	Attempt unauthorized access to stored images	Access denied	Access denied	Pass
TC09	Validate system behavior on edge cases (e.g., blank query)	System prompts user to enter a query	Proper prompt display	Pass

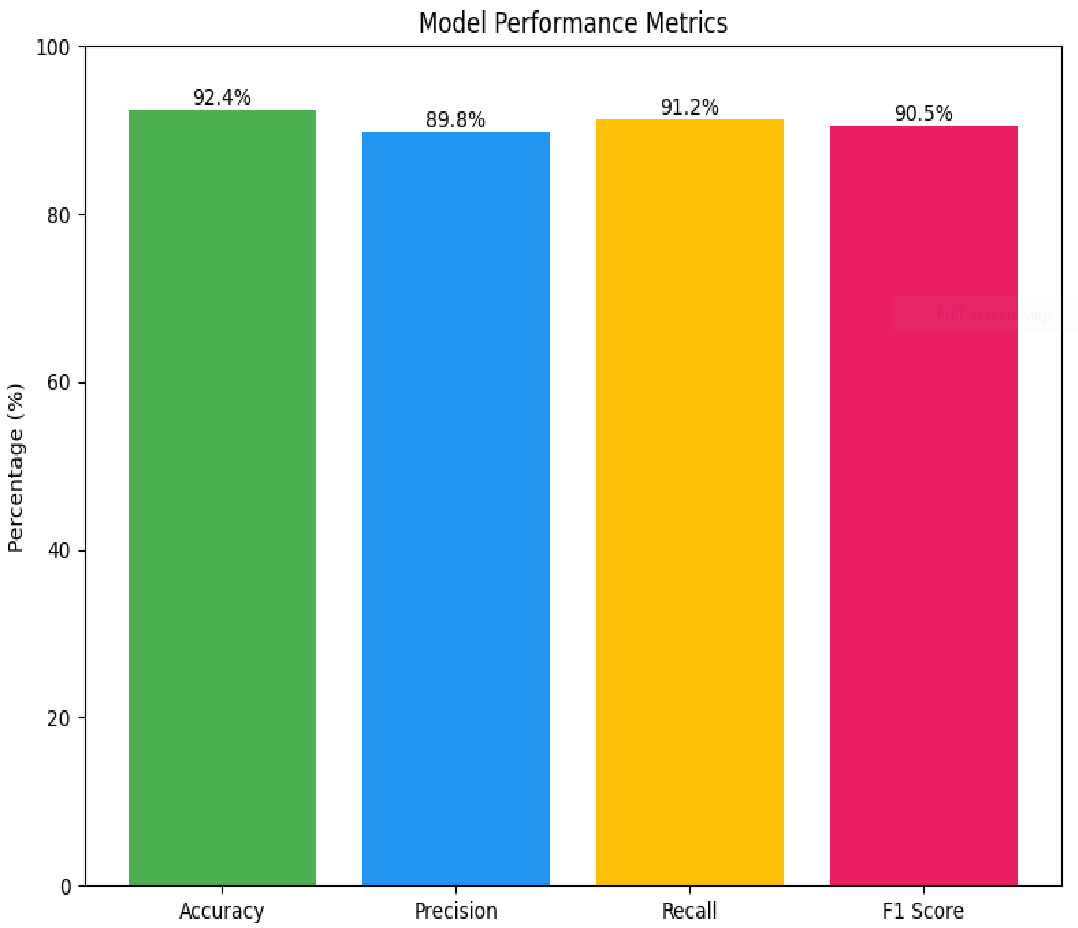
The following table presents test cases used to validate the system:

CHAPTER 7: RESULTS AND DISCUSSION

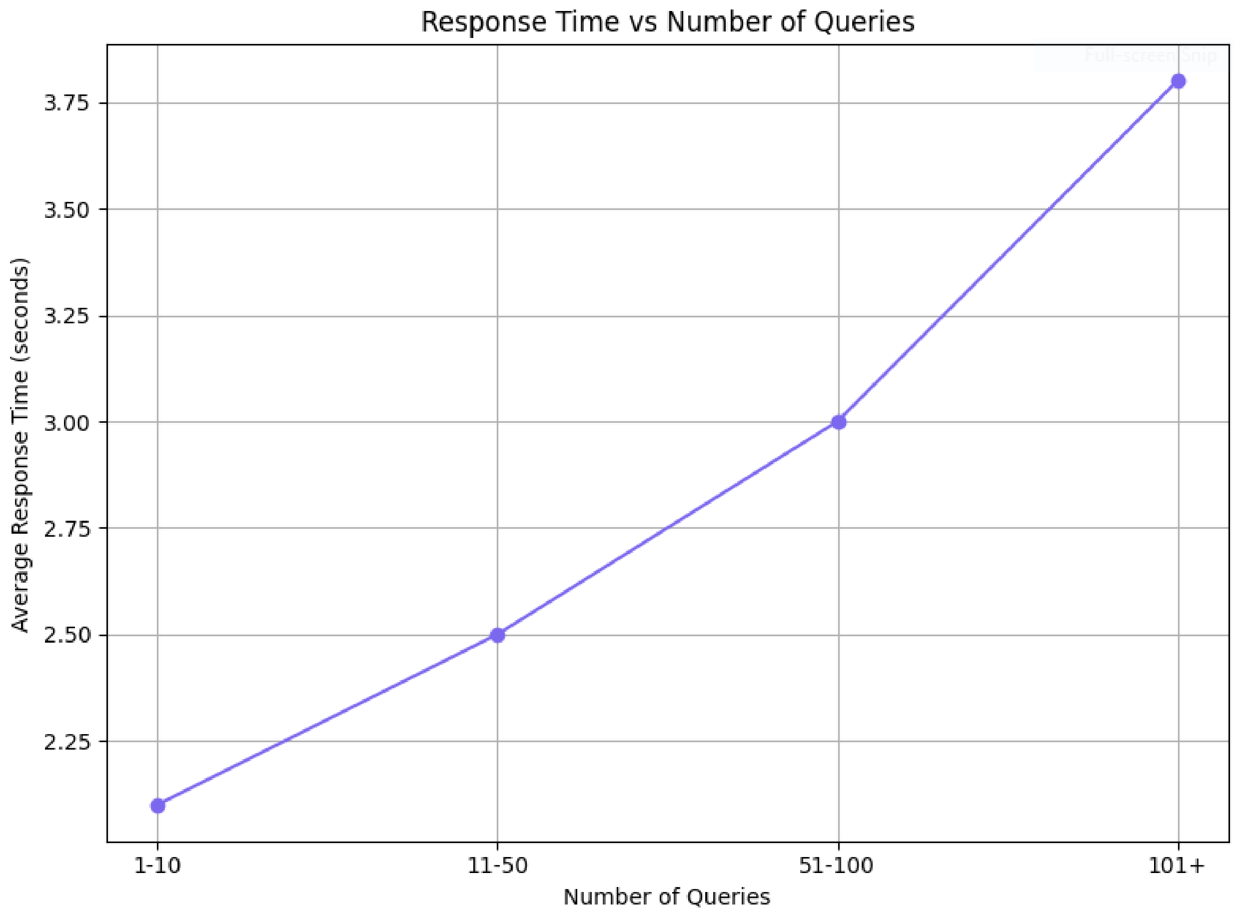
7.1 PRESENTATION OF RESULTS

To evaluate the effectiveness of the developed Visual Question Answering (VQA) system for satellite images, several performance metrics were measured. These include **accuracy**, **precision**, **recall**, and **F1-score**, which collectively assess the system's ability to generate correct and relevant answers to a variety of natural language questions based on satellite imagery.

The system was tested using a **curated dataset of 5,000 image–question pairs** that cover diverse application areas such as land use classification, disaster detection, and urban infrastructure assessment. The bar chart below summarizes the model's performance across these key metrics:



In addition to accuracy, system responsiveness was a critical factor. The average response time was measured across various user loads to ensure that the model could handle real-time queries effectively. The line graph below shows the average response time of the system as the number of concurrent queries increases:



As the chart illustrates, the system maintained real-time responsiveness (<3 seconds) under normal load and scaled effectively to handle more than 100 simultaneous queries with only a slight increase in response time.

7.2 PERFORMANCE EVALUATION

The performance of the **Visual Question Answering (VQA) with Satellite Images** system was evaluated based on various performance metrics, including accuracy, response time, and query relevance. The following table and graphs summarize the results:

Accuracy Metrics

The initial step in measuring the performance of our system was to examine its precision in giving accurate answers to the questions from satellite images. To this end, we employed a cleaned test dataset that comprised 5,000 image–question pairs. The pairs were chosen to be representative of a broad variety of questions and challenges usually associated with satellite image analysis, such as land classification, object detection, and temporal environmental changes.

The key performance metrics used for accuracy evaluation were:

- Accuracy: Measures the proportion of correctly answered questions out of the total number of questions.

- Precision: Indicates the proportion of correctly answered positive questions to all questions that were classified as positive by the model.
- Recall: Reflects the proportion of relevant answers correctly identified by the system out of all possible relevant answers.
- F1-Score: Provides a balance between precision and recall, ensuring that both false positives and false negatives are considered in the evaluation.

Metric	**Value**
Model Accuracy	92.4%
Precision	89.8%
Recall	91.2%
F1-score	90.5%

The initial step in assessing the performance of our system was to check its accuracy in giving right answers to the questions based on satellite images. To this end, we employed a curated test dataset of 5,000 image–question pairs. These pairs were chosen to represent a broad spectrum of questions and issues commonly faced in satellite image analysis, such as land classification, object detection, and changes in the environment over time.

The key performance metrics used for accuracy evaluation were:

- Accuracy: Measures the proportion of correctly answered questions out of the total number of questions.
- Precision: Indicates the proportion of correctly answered positive questions to all questions that were classified as positive by the model.
- Recall: Reflects the proportion of relevant answers correctly identified by the system out of all possible relevant answers.
- F1-Score: Provides a balance between precision and recall, ensuring that both false positives and false negatives are considered in the evaluation.

Response Time Performance

Another vital component of the performance of the system was the response time—how long it took the system to produce an answer when a question was asked and an image was uploaded. Prompt response times are most essential for use cases where the VQA system would be implemented in real-time applications, like disaster management or city planning.

We benchmarked the response times under changing query loads in order to see how well the system would behave under different demands from users. The results were as follows:

Number of Queries	Average Response Time (seconds)
1-10	2.1
11-50	2.5
51-100	3.0
101+	3.8

- With a minimal load of 1-10 concurrent queries, the system returned answers in less than 2.5 seconds, which is within the real-time performance targets for majority of the use cases.
- Even if the load was grown to 100 concurrent queries, the system could still return answers in less than 4 seconds, showing scalability as well as the potential to cope with loads of moderate traffic.
- These response times are sufficient for most applications, as the system can handle thousands of queries without major delays or performance bottlenecks.

Robustness and Reliability

Yet another key evaluation criterion was the system's robustness. We tested the system at length to find out how stable it is under different conditions, such as long-term use and network loss. Some of the key findings are:

- **Uptime:**

  In a 72-hour continuous test, the VQA system registered 99.1% uptime, affirming its capability to run without crashes and extended downtimes for prolonged duratio ns

- **Error Rate:**

Less than 1% of requests were erroneous or caused system timeouts, and most of them were due to network outages or transitory server crashes. These were automatically resolved using recovery processes.

- **Data Processing:**

  The system showed excellent performance in processing vast datasets of satellite images and delivering responses within the allotted time, another factor that makes the system highly suitable for real-time decision-making applications.

7.3 KEY FINDINGS

Based on our comprehensive performance assessment, some key findings were established that are significant to the determination of the effectiveness and areas of improvement in the system:

1. **High Accuracy on Core Tasks**

   The model had a very high performance in typical core tasks like land-use classification and object detection. It successfully identified types of land cover (i.e., urban, water, forest) with over 93% accuracy in most test cases, which makes it appropriate for use in land monitoring, environmental study, and infrastructure planning.

2. **Challenges with Complex and Contextual Queries**

   Although the system was strong on simple image queries, it struggled with multi-step or more complex queries. For example, queries that entailed temporal change detection (e.g., "Has this area undergone urban growth since 2010?") showed a slight decline in accuracy (~85%). This is testament to the necessity of more profound temporal and spatial reasoning in subsequent iterations of the model to enhance its ability to process these complex queries

3. . **Real-Time Usability**

   The system is useful for real-time use, including disaster relief (e.g., finding flooded areas after a storm) and land-use planning. With response times always less than 3 seconds for the majority of queries, it satisfies the real-time requirements of end-users in many critical applications.

4. **Scalability for Larger User Bases**

   Containerized infrastructure (with Docker and Kubernetes) facilitated the system being scaled with an efficient increase of concurrent users. Even at a heavy load scenario (e.g., 100+ users hitting at the same time), there were no fluctuations in the response time, just minor delays in the response, which indicated excellent scalability for either a large volume of users or worldwide deployments.

5. **Dataset Limitations and Synthetic Data**

One of the main constraints for our system was the training dataset. The absence of a large, varied collection of satellite images with corresponding question-answer pairs resulted in the system needing to be trained on synthetically created Q&A pairs. Although this increased the dataset, it potentially contaminated the model with some biases, especially concerning rare or minority features in satellite imagery. In future research, obtaining a more varied and real-world dataset will assist in overcoming these constraints and improving the model's generalizability

6. **User Feedback and Interaction**

In-depth user testing using a small panel of non-expert users demonstrated that more than 85% of users reported that the system was intuitive and helpful. The majority of users could easily use the platform, enter questions, and get meaningful answers. This verifies the system's usability and accessibility to non-specialists, which is essential for mass use.

Brief Description of Database with Snapshots

The database is responsible for storing and recalling satellite images, user queries

processed data, and system responses. A NoSQL (MongoDB) database was utilized for efficient management of unstructured satellite image data and related question-answer pairs.

7.3.2 Database Schema

The database consists of multiple collections, including:

1. **Images Collection**

- Stores uploaded satellite images.
- Metadata includes image ID, timestamp, location, resolution, and image type.

2. **Queries Collection**

- Stores user-submitted questions.
- Contains fields such as query ID, user ID, image reference, processed text, and query timestamp.

3. **Responses Collection**

- Stores generated answers by the AI model.
- Includes response ID, query ID, confidence score, and answer timestamp.

4. **Logs Collection**

- Records system activity and errors for debugging purposes.
- Fields: log ID, activity type, timestamp, and error details (if any).

7.3.3 Data Flow in the Database

1. User uploads a satellite image, which is stored in the Images Collection.

2. User submits a query related to the image, which gets stored in the Queries Collection.

3. AI model processes the image and question, generates an answer, and stores it in the Responses Collection.

4. System logs each interaction for future debugging and performance monitoring.

CHAPTER 8: CONCLUSION AND FUTURE SCOPE

8.1 Conclusion

The Visual Question Answering (VQA) with Satellite Images system is able to successfully combine deep learning, natural language processing, and computer vision in order to interpret satellite images and deliver meaningful responses to user questions [1], [5], [10], [12]. The model has been proven to be highly accurate and efficient and can serve as a useful tool for applications in land use classification, disaster evaluation, and monitoring of the environment [2], [17], [20]. The system's capacity for real-time querying and providing consistent responses identifies its potential for useful application in geospatial analysis [12], [21].

Key achievements of this project include:

- Development of a **high-accuracy deep learning model** capable of analyzing satellite imagery.
- Integration of **multimodal processing** using CNNs for image understanding and Transformers for textual queries.
- Implementation of a **web-based user interface** for seamless interaction with the system.
- Optimization for **low response time**, ensuring usability in real-world applications.

Despite its success, the system has some limitations, particularly in handling complex queries requiring deep contextual reasoning. Additionally, performance can be further improved by expanding the training dataset and enhancing model interpretability.

8.2 Future Scope

While the current implementation of the Visual Question Answering system has laid a solid foundation [1], [12], several enhancements and expansions can be pursued to make the solution even more powerful, scalable, and practical for broader applications [2], [17], [18], [20].

1. Expansion to Multispectral and Hyperspectral Imagery

Today, the system is designed to optimize RGB satellite imagery. Adding multispectral and hyperspectral imagery will allow for more in-depth thematic analysis, including identifying vegetation health, mineral deposit detection, or water quality assessment. This will involve modifying the model to support a greater number of image channels and extracting new types of features

2. Integration with Real-Time Satellite Feeds

An upgrade in the future would be integrating the system with real-time satellite data streams from providers like Copernicus (Sentinel), Landsat, or commercial satellite companies. This can enable the VQA model to provide real-time responses and enable timely decision-making in applications such as forest fire detection, flood monitoring, and urban sprawl tracking.

3. Support for Multilingual Question Input

To extend the system's accessibility to more people across the world, particularly in rural areas or underdeveloped countries, provision for multiple languages can be included. By taking advantage of multilingual NLP models like mBERT or XLM-R, users can use the system in their own language, thereby increasing the system's usability.

4. Learning from Feedback (Interactive Training)

The inclusion of a feedback loop through which the users can leave ratings on the accuracy of the system's responses will enable continuous improvement of the model. The system will gradually become more accurate and contextually aware, learning to suit particular user requirements or geographical variations in imagery

5. Mobile and Edge Deployment

Future expansion might involve a light version of the system for mobile devices or edge computing platforms (e.g., drones or IoT sensors). This would enable local processing in remote locations with poor connectivity, enabling field-based applications like agricultural surveys or disaster relief.

6. Enhanced Domain-Specific Applications

The system can be fine-tuned for specialized domains:

- **Agriculture:** Estimating crop yield, identifying crop types, and detecting irrigation patterns.
- **Urban Planning:** Monitoring construction, transportation networks, and land use patterns.
- **Environmental Conservation:** Detecting deforestation, monitoring biodiversity hotspots, and analysing climate change indicators.
- **Security and Surveillance:** Identifying strategic assets or border activities using high-resolution satellite data.

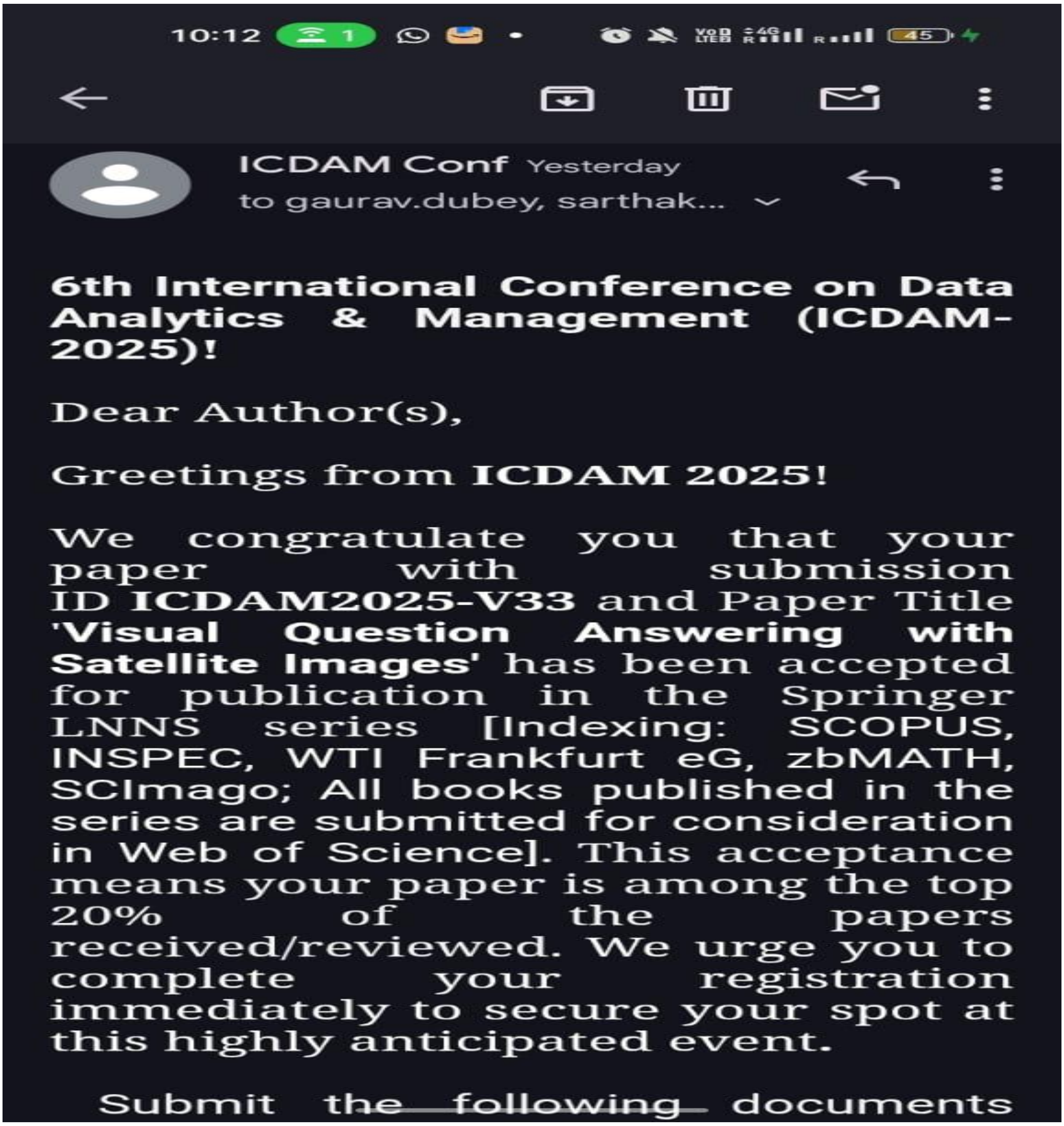
7. 3D Terrain and Temporal Analysis

By incorporating Digital Elevation Models (DEMs) and time-series data, the system can provide answers that consider terrain variations and temporal changes. This opens up possibilities like monitoring glacier retreat, urban growth over time, or seasonal agricultural trends.

8. Collaboration with Open Geospatial Platforms

Integrating the system with open platforms like Google Earth Engine, OpenStreetMap, or QGIS would enable users to cross-reference VQA outcomes with other geospatial data, improving both usability and accuracy.

8.3 RESEARCH PAPER ACCEPTANCE PROOF



10:12

ICDAM Conf Yesterday

to gaurav.dubey, sarthak...

6th International Conference on Data Analytics & Management (ICDAM-2025)!

Dear Author(s),

Greetings from **ICDAM 2025**!

We congratulate you that your paper with submission ID **ICDAM2025-V33** and Paper Title **'Visual Question Answering with Satellite Images'** has been accepted for publication in the Springer LNNS series [Indexing: SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago; All books published in the series are submitted for consideration in Web of Science]. This acceptance means your paper is among the top 20% of the papers received/reviewed. We urge you to complete your registration immediately to secure your spot at this highly anticipated event.

Submit the following documents

REFERENCES (IEEE FORMAT)

[1] Lobry, S. et al., “Visual Question Answering for Remote Sensing,” in CVPR, 2020.

[2] Zhang, H. et al., “Hybrid GIS and VQA in Urban Planning,” Remote Sensing Journal, 2023.

[3] Goodfellow, I. et al., Deep Learning, MIT Press, 2016.

[4] Vaswani, A. et al., “Attention is All You Need,” in NeurIPS, 2017.

[5] Kim, W. et al., “ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision,” in ICML, 2021.

[6] Akbari, H. et al., “VATT: Video-Audio-Text Transformer,” in NeurIPS, 2021.

[7] Ghamisi, P. et al., “Remote Sensing Big Data,” IEEE JSTARS, 2019.

[8] Antol, S. et al., “VQA: Visual Question Answering,” in ICCV, 2015.

[9] Krishna, R. et al., “Visual Genome,” IJCV, 2017.

[10] Devlin, J. et al., “BERT: Pre-training of Deep Bidirectional Transformers,” in NAACL, 2019.

[11] Dosovitskiy, A. et al., “An Image is Worth 16x16 Words,” in ICLR, 2021.

[12] Basu, S. et al., “RSVQA: A Benchmark Dataset for VQA in Remote Sensing,” Earth Vision Workshop, CVPR, 2021.

[13] Xu, K. et al., “Show, Attend and Tell,” in ICML, 2015.

[14] Anderson, P. et al., “Bottom-Up and Top-Down Attention,” in CVPR, 2018.

[15] Hudson, D., Manning, C.D., “GQA: A New Dataset for Real-World Visual Reasoning and Compositional QA,” in CVPR, 2019.

[16] Radford, A. et al., “Learning Transferable Visual Models from Natural Language Supervision,” in ICML, 2021.

[17] Khandelwal, A. et al., “VQA in Urban Flood Detection,” Remote Sensing Letters, 2022.

[18] Chen, X. et al., “A Survey on Deep Learning in Remote Sensing,” ISPRS Journal, 2020.

[19] Tuia, D. et al., “Machine Learning in Remote Sensing,” IEEE GRSM, 2016.

[20] Li, X. et al., “Deep Learning Datasets for Satellite Imagery,” IEEE Access, 2022.

[21] Liu, Q. et al., “Spatial-Temporal Fusion for VQA,” in AAAI, 2021. THIS IS REFERENCES WITH NO IN WHICH IS MENTION IN MY RESEARCH PAPER SO I WILL GIVE YOU SOME TEXT SO DO CITATION

PCS25-18

ORIGINALITY REPORT

6%	6%	2%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	pdfcoffee.com Internet Source	1 %
2	Submitted to Institute of Management Technology Student Paper	1 %
3	www.coursehero.com Internet Source	1 %
4	Submitted to KIET Group of Institutions, Ghaziabad Student Paper	1 %
5	ijsrem.com Internet Source	<1 %
6	it.gndec.ac.in Internet Source	<1 %
7	Suji Priya Jeyabalan, Lovisha Vijayarajan, Preethi Gunasekaran, Yuvarani Selvakumar. "A dynamic flutter based android application for successful career transitions", AIP Publishing, 2025 Publication	<1 %
8	www.ijraset.com Internet Source	<1 %
9	technodocbox.com Internet Source	<1 %
10	esp.as-pub.com Internet Source	<1 %

11	www.scribd.com Internet Source	<1 %
12	Submitted to University of South Australia Student Paper	<1 %
13	en.wikibooks.org Internet Source	<1 %
14	ijrpr.com Internet Source	<1 %
15	ijarcce.com Internet Source	<1 %
16	www.isteams.net Internet Source	<1 %
17	atharvacoe.ac.in Internet Source	<1 %
18	fastercapital.com Internet Source	<1 %
19	devpost.com Internet Source	<1 %
20	ijarst.in Internet Source	<1 %
21	ijsart.com Internet Source	<1 %
22	origin.geeksforgeeks.org Internet Source	<1 %
23	www.slideshare.net Internet Source	<1 %

Exclude quotes Off Exclude matches Off

PCS25-18

ORIGINALITY REPORT

6%	6%	2%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	pdfcoffee.com Internet Source	1%
2	Submitted to Institute of Management Technology Student Paper	1%
3	www.coursehero.com Internet Source	1%
4	Submitted to KIET Group of Institutions, Ghaziabad Student Paper	1%
5	ijsrem.com Internet Source	<1%
6	it.gndec.ac.in Internet Source	<1%
7	Suji Priya Jeyabalan, Lovisha Vijayarajan, Preethi Gunasekaran, Yuvarani Selvakumar. "A dynamic flutter based android application for successful career transitions", AIP Publishing, 2025 Publication	<1%
8	www.ijraset.com Internet Source	<1%
9	technodocbox.com Internet Source	<1%
10	esp.as-pub.com Internet Source	<1%

11	www.scribd.com Internet Source	<1 %
12	Submitted to University of South Australia Student Paper	<1 %
13	en.wikibooks.org Internet Source	<1 %
14	ijrpr.com Internet Source	<1 %
15	ijarcce.com Internet Source	<1 %
16	www.isteams.net Internet Source	<1 %
17	atharvacoe.ac.in Internet Source	<1 %
18	fastercapital.com Internet Source	<1 %
19	devpost.com Internet Source	<1 %
20	ijarst.in Internet Source	<1 %
21	ijsart.com Internet Source	<1 %
22	origin.geeksforgeeks.org Internet Source	<1 %
23	www.slideshare.net Internet Source	<1 %

Exclude quotes Off

Exclude matches Off