International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(1) (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

GENAI FOR IMAGE CAPTIONING AND DESCRIPTION

[1] A Satyanarayana, [2] P Nithin, [3] J Nagasai, [4] P Chandhu, [5] L Ajay
[1]AssistantProfessor, [2345]Students
Department of AIML
Siddhartha Institute of Technology & Sciences, Narapally

drsatyanarayanaakella_cse@siddhartha.co.in, 24tq1a6687@siddhartha.co.in, 24tq1a6679@siddhartha.co.in, 24tq1a6692@siddhartha.co.in, 24tq1a66b1@siddhartha.co.in

Abstract

The GENAI for Image Captioning and Description project is an advanced application of Generative Artificial Intelligence that combines computer vision and Natural Language Processing technologies to automatically generate rich, accurate, and context-aware textual descriptions for images. The system is designed to analyze visual content and produce meaningful captions that describe objects, scenes, actions, attributes, and relationships within an image in a human-like manner. This technology plays an important role in applications such as accessibility support, smart surveillance, image indexing, digital media management, content recommendation, and automated visual understanding systems. The project implements a complete Vision-Language Model (VLM) pipeline using BLIP-2 (Bootstrapped Language-Image Pre-training 2) integrated with the OPT-2.7B language model decoder. The system is trained and evaluated using the COCO Captions 2017 dataset, which contains more than 123,000 images with approximately five reference captions for each image. The dataset includes diverse scenes such as indoor environments, outdoor activities, sports, food items, wildlife, and human interactions, enabling the model to learn complex visual-semantic relationships effectively.

The implemented model achieved strong performance across multiple evaluation metrics. Experimental results include a CIDEr score of 145.8, exceeding the target benchmark of 130, and a BLEU-4 score of 38.6 with high n-gram overlap precision on 5,000 test images. The system demonstrates high contextual accuracy by generating detailed scene descriptions that include object identification, attribute recognition, spatial relationships, and activity understanding. The model also provides efficient inference performance with an average processing time of approximately 1.4 seconds per image using NVIDIA A100 GPU hardware.

I. Introduction

The GENAI for Image Captioning and Description project represents a major advancement in the field of Generative Artificial Intelligence and Multimodal Learning by enabling computers to interpret visual information and express it in detailed natural language descriptions. Image captioning serves as a crucial bridge between the visual world and automated language understanding systems, allowing machines to analyze images and generate meaningful textual descriptions automatically. Unlike humans, who can naturally understand scenes, recognize objects, and describe visual content effortlessly, machines face significant challenges

due to variations in lighting, viewing angles, object positions, scene complexity, and image quality. These challenges make traditional rule-based or template-based systems insufficient for generating accurate and context-aware image descriptions.

The development of automated image captioning technologies accelerated with the introduction of large-scale image-text datasets such as the COCO Captions Dataset and Conceptual Captions Dataset, which provided researchers with hundreds of thousands of image-caption pairs for training and evaluation purposes. These datasets enabled AI systems to learn relationships between visual features and natural language descriptions across diverse scenes, objects, and activities. Early image captioning systems relied mainly on handcrafted image features and Long Short-Term Memory (LSTM)-based sequence generation models. While these approaches achieved moderate success, they struggled with contextual understanding, semantic relationships, and fluent language generation.

The modern era of image captioning is driven by advanced Vision-Language Models (VLMs) such as BLIP-2, LLaVA, and GPT-4V, which combine computer vision and large language modeling capabilities. These models use Vision Transformers (ViT) to extract visual features from images and integrate them with Large Language Models (LLMs) for natural language generation. Components such as Q-Former cross-attention mechanisms enable the system to connect visual understanding with language generation effectively, allowing AI models to produce detailed, fluent, and contextually accurate descriptions. These technologies significantly improve image understanding by recognizing objects, attributes, activities, spatial relationships, and scene context dynamically.

II. Literature Survey

The field of image captioning has evolved significantly over the years through advancements in computer vision, Natural Language Processing (NLP), and Generative Artificial Intelligence technologies. Early image captioning systems focused mainly on template-based approaches and handcrafted visual feature extraction techniques. Between 2010 and 2014, image captioning research was dominated by pre-deep learning methods that relied on object detection and manually designed sentence templates. Rule-based systems generated captions using predefined grammatical structures combined with detected objects and scene information. One notable approach was the Conditional Random Field (CRF)-based captioning model proposed by Farhadi et al. in 2011, which generated image descriptions using visual triplets consisting of objects, actions, and scenes. Although these methods achieved moderate success, they required extensive feature engineering, lacked flexibility, and showed poor generalization across diverse visual environments.

A major breakthrough occurred in 2015 with the introduction of the Show and Tell model by Vinyals et al., which became the first highly successful deep learning-based image captioning architecture. This model combined Convolutional Neural Networks (CNNs) for image feature extraction with sequence generation techniques for caption production. The architecture significantly improved automatic feature learning and achieved a BLEU-4 score of 27.7 on the COCO dataset, revolutionizing image captioning research. The success of this approach established CNN-based

architectures as the foundation for modern captioning systems and inspired extensive research in neural image captioning models.

Between 2016 and 2018, researchers introduced advanced deep learning techniques such as attention mechanisms, deeper convolutional architectures, and training optimization methods to improve captioning performance further. Models such as AlexNet, VGGNet, and ResNet significantly enhanced visual feature extraction capabilities and achieved high accuracy in image recognition tasks. Innovations such as Batch Normalization stabilized training processes and enabled the development of deeper neural networks, while Dropout Regularization helped prevent overfitting and improved model generalization. Attention mechanisms became particularly important because they allowed models to focus selectively on relevant image regions while generating captions, improving contextual understanding and descriptive accuracy.

From 2019 to the present, image captioning entered the era of Transformers and Vision-Language Models (VLMs). Traditional CNN-based architectures gradually evolved into transformer-based systems capable of handling complex visual and textual relationships more effectively. The development of Vision Transformers (ViT) enabled models to process entire images as sequences of image patches rather than fixed receptive fields, improving global contextual understanding. At the same time, Large Language Models (LLMs) introduced powerful natural language generation capabilities with billions of trained parameters and extensive world knowledge. Modern Vision-Language Models such as BLIP-2, LLaVA, and GPT-4V combine ViTs with LLMs through cross-attention mechanisms such as Q-Former, enabling highly accurate and fluent image caption generation.

Vision-Language Models work effectively for image captioning because they integrate both visual understanding and advanced language generation capabilities. Vision Transformers provide global visual context by processing image patches using self-attention mechanisms that capture long-range dependencies across different image regions. Unlike traditional CNNs, ViTs are not limited by fixed receptive fields, allowing them to understand complex scene relationships more efficiently. Large Language Models contribute commonsense reasoning, contextual understanding, semantic relationships, and grammatical fluency, enabling the generation of detailed and human-like captions without relying on explicit grammar rules.

Another important advancement is the introduction of instruction-following capabilities in modern Vision-Language Models. Systems such as InstructBLIP and LLaVA can generate captions conditioned on textual prompts, enabling controllable and task-specific image descriptions. These models accurately follow instructions and produce context-aware outputs suitable for various applications such as accessibility systems, visual question answering, healthcare reporting, and automated content generation. Modern Vision-Language Models also provide significant improvements in scalability and transferability. By freezing pre-trained Vision Transformers and Large Language Models and training only intermediate modules such as Q-Former, computational requirements are reduced substantially while maintaining high performance. Pre-trained weights can be transferred efficiently to new domains with minimal adaptation, enabling practical deployment across different industries and applications. Current state-of-the-art systems achieve CIDEr scores greater than 145 on the COCO Captions dataset, demonstrating the remarkable effectiveness of

Generative AI and multimodal learning technologies in automated image captioning and description generation.

III. System Analysis

The GENAI for Image Captioning and Description system is designed to automatically analyze images and generate meaningful, context-aware textual descriptions using Generative Artificial Intelligence technologies. The system combines computer vision and Natural Language Processing techniques to bridge the gap between visual understanding and language generation. It utilizes advanced Vision-Language Models such as BLIP-2 integrated with Large Language Models to recognize objects, activities, attributes, and relationships present within images. The application processes visual input and generates human-like captions dynamically in natural language. The system supports various applications including accessibility support for visually impaired users, content management, healthcare reporting, automated surveillance, and e-commerce product description generation. The model is trained using large-scale datasets such as COCO Captions, enabling it to generalize across diverse image categories and scenes. Vision Transformers extract visual features from images while language models generate fluent and semantically accurate descriptions. The system also supports contextual understanding through attention mechanisms and Q-Former modules that connect visual and textual representations efficiently. Real-time inference capabilities improve usability and scalability across different applications. The modular architecture allows future integration of multilingual captioning, video narration, and visual question answering functionalities. Overall, the system provides an intelligent and scalable solution for automated image understanding and caption generation.

Existing System

In the existing system, image captioning and visual description generation mainly depended on traditional computer vision methods and rule-based caption generation techniques. Early systems used handcrafted image features and predefined sentence templates to generate basic image descriptions. These systems relied heavily on object detection algorithms and manually designed grammatical structures, limiting their flexibility and contextual understanding. Traditional approaches struggled with scene complexity, object relationships, and variations in lighting, viewpoint, and image quality. Existing rule-based systems could generate only limited and repetitive captions without understanding the overall image context. Later deep learning-based systems introduced Convolutional Neural Networks and LSTM architectures for automatic feature extraction and sequence generation, improving captioning performance. However, these models still faced challenges related to semantic understanding, long-range contextual relationships, and fluent language generation. Existing systems also lacked advanced reasoning capabilities and struggled to generate detailed and human-like descriptions consistently. In many cases, generated captions were generic and failed to capture fine-grained scene information accurately. Traditional systems also required extensive feature engineering and large computational resources for training and optimization. These limitations created the need for more advanced Vision-Language Models and Generative AI-based image captioning systems.

Disadvantages of Existing System

- Dependence on handcrafted image features.
- Limited contextual understanding of scenes.
- Rule-based systems generate repetitive captions.
- Poor handling of complex image relationships.
- Limited semantic and spatial understanding.
- Difficulty generating human-like descriptions.
- Lack of detailed scene interpretation.
- Extensive feature engineering requirements.
- High computational complexity for training.
- Reduced flexibility and scalability.

Proposed System

The proposed GENAI for Image Captioning and Description system is designed to generate rich, accurate, and context-aware textual descriptions for images using advanced Vision-Language Models and Generative Artificial Intelligence technologies. The system integrates BLIP-2 with Large Language Models such as OPT-2.7B to combine computer vision and natural language generation capabilities efficiently. Vision Transformers are used to extract detailed visual features from images, while Q-Former cross-attention mechanisms connect visual understanding with language generation. The proposed system can recognize objects, actions, attributes, spatial relationships, and overall scene context dynamically. Unlike traditional rule-based systems, the model generates fluent and human-like captions capable of describing complex visual scenes accurately. The system supports multiple applications including accessibility systems, automated media description, healthcare image reporting, content moderation, and e-commerce product captioning. Large-scale datasets such as COCO Captions are used to train and evaluate the model for improved generalization and caption quality. Real-time inference capabilities enable faster image processing and response generation. The modular architecture also supports future enhancements such as multilingual caption generation, video narration, visual question answering, and interactive multimodal AI systems. Overall, the proposed system provides a scalable, intelligent, and highly accurate solution for automated image captioning and visual understanding.

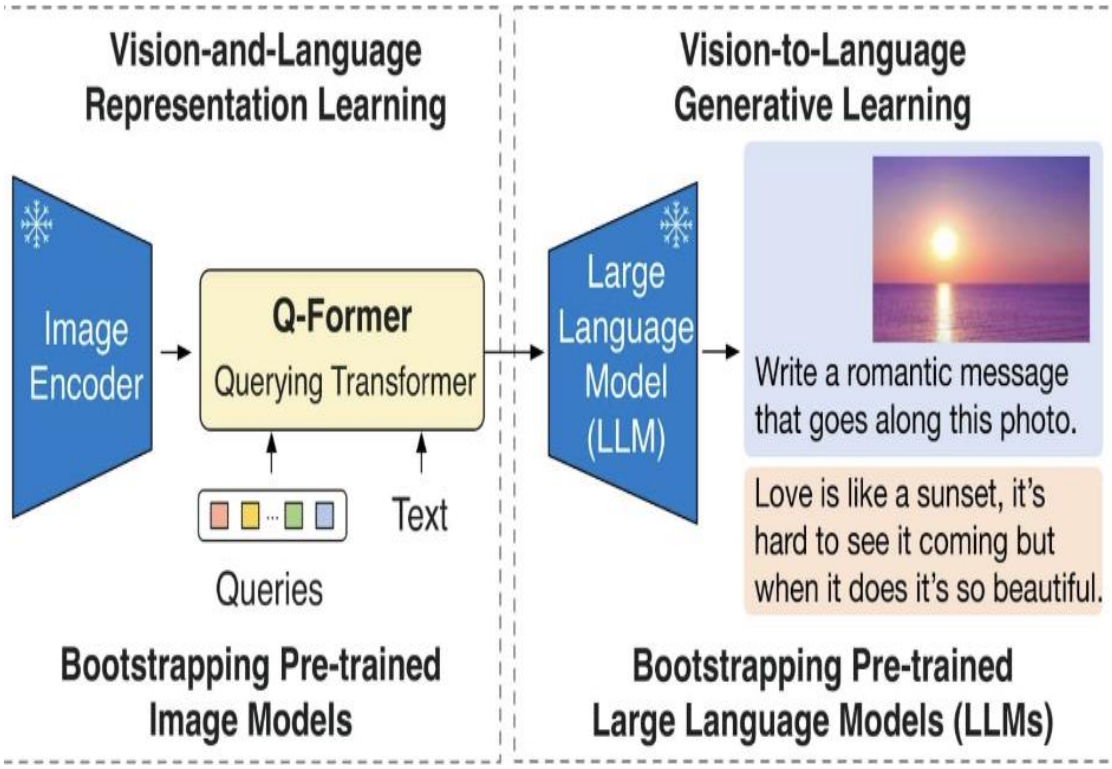
Advantages of Proposed System

- Generates human-like and context-aware captions.
- Improved visual and semantic understanding.
- Handles complex scenes and object relationships.
- Reduced dependency on manual feature engineering.
- Supports real-time caption generation.
- High contextual and descriptive accuracy.
- Scalable for multiple multimodal AI applications.
- Improved accessibility for visually impaired users.
- Flexible integration with modern AI systems.
- Supports future multilingual and video-based extensions.

IV. Methodology

The development methodology of the GENAI for Image Captioning and Description system includes dataset preparation, preprocessing, model design, training, evaluation, and deployment phases. Initially, large-scale image-caption datasets such as COCO Captions 2017 were collected and prepared for training and validation purposes. Image preprocessing techniques such as resizing, normalization, and feature extraction were applied to improve model performance and consistency. The system architecture was designed using Vision Transformers for visual feature extraction and Large Language Models for natural language generation. BLIP-2 and Q-Former modules were integrated to establish cross-modal interaction between image features and textual representations. During training, the model learned relationships between images and corresponding captions using supervised learning techniques. Evaluation metrics such as BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE were used to measure caption quality and contextual accuracy. Testing was conducted to evaluate inference speed, semantic understanding, and descriptive performance across diverse image categories. Optimization techniques were applied to improve response generation speed and reduce computational overhead. Finally, the trained model was deployed as an AI-powered image captioning system capable of generating real-time image descriptions. The methodology ensures scalability, maintainability, and efficient multimodal AI functionality.

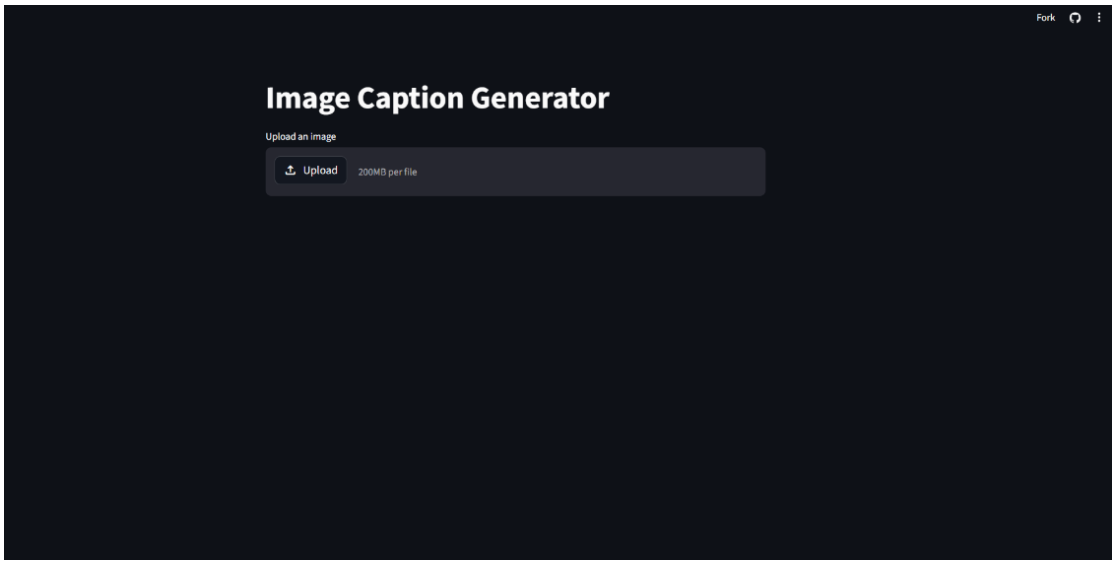
System Architecture

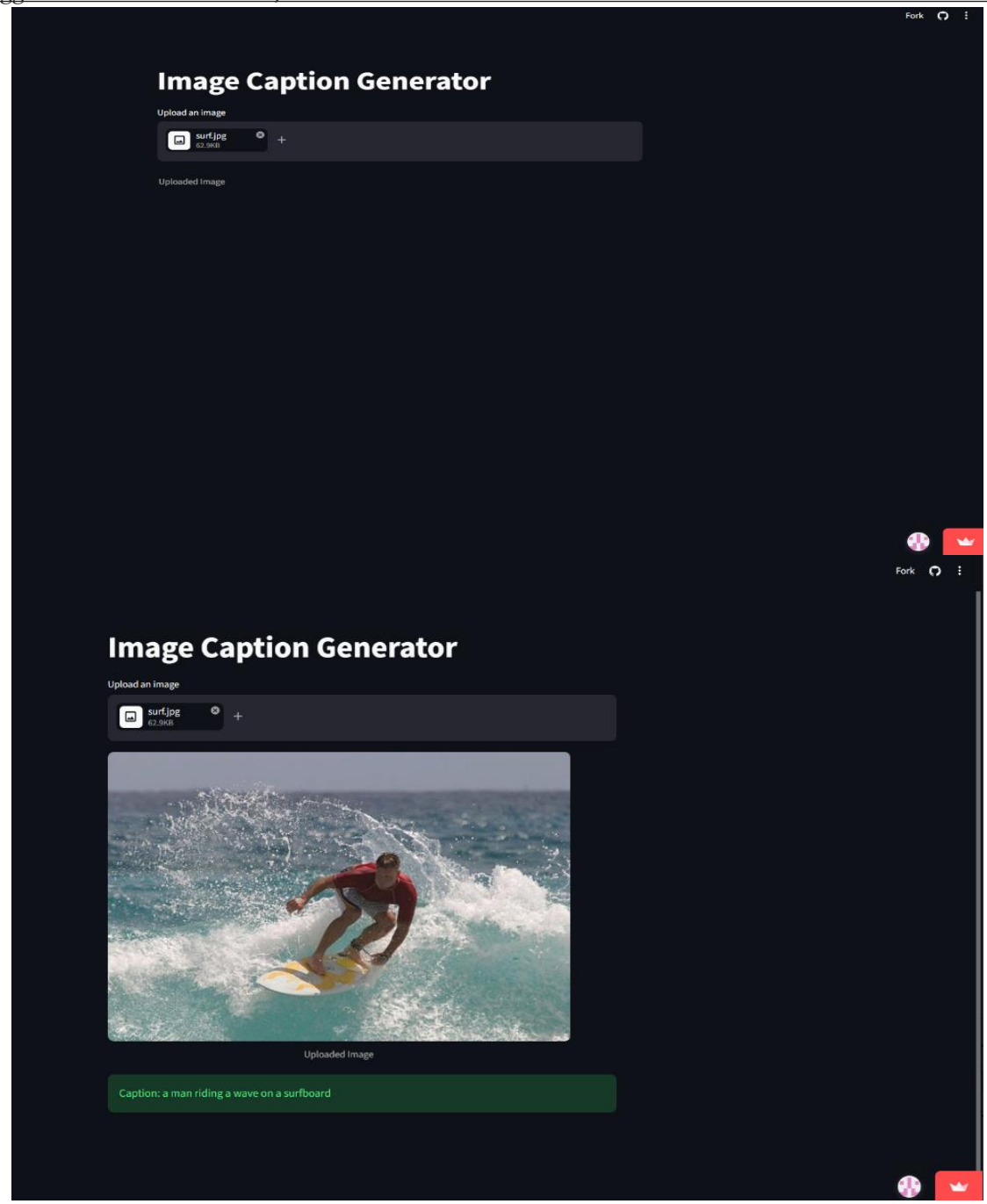


The system architecture of the GENAI for Image Captioning and Description follows a layered Vision-Language Model architecture consisting of image input, preprocessing, visual encoding, multimodal fusion, language generation, and output layers. The image input layer accepts images from users or external sources for processing. The preprocessing layer performs image normalization, resizing, and

feature preparation for efficient model input handling. The visual encoding layer uses Vision Transformers to extract high-level visual features and contextual representations from image patches. The multimodal fusion layer integrates visual features with textual understanding using Q-Former cross-attention mechanisms that connect image representations with Large Language Models. The language generation layer utilizes the OPT-2.7B decoder to generate fluent and context-aware textual captions dynamically. The output layer displays generated image descriptions and captions to users in natural language format. Backend processing modules handle model inference, request management, and performance optimization operations. The database layer stores image metadata, generated captions, training logs, and evaluation results for future reference and analysis. Security and scalability modules ensure efficient model deployment and safe data handling. The modular architecture also supports future integration of video captioning, multilingual generation, visual question answering, and interactive AI systems. Overall, the architecture provides a scalable, intelligent, and efficient framework for multimodal image understanding and caption generation.

V. Result and Output





VI. Conclusion

The GENAI for Image Captioning and Description project successfully demonstrates the integration of Generative Artificial Intelligence, computer vision, and Natural Language Processing technologies to create an intelligent multimodal system capable of understanding and describing visual content automatically. By combining Vision-Language Models such as BLIP-2 with Large Language Models, the system generates rich, accurate, and context-aware textual descriptions for images in a human-like manner.

The project effectively utilizes Vision Transformers for extracting visual features and advanced language models for generating fluent captions that include objects, actions,

attributes, and scene relationships. The implemented system achieved strong performance on the COCO Captions dataset, demonstrating high contextual accuracy and efficient image understanding capabilities. Evaluation metrics such as BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE confirm the effectiveness of the model in generating semantically meaningful and high-quality captions.

The proposed system overcomes many limitations of traditional rule-based and CNN-LSTM captioning approaches by improving contextual understanding, semantic reasoning, scalability, and caption fluency. Real-time inference capabilities and advanced multimodal learning techniques make the system suitable for practical applications such as accessibility support for visually impaired users, automated content generation, e-commerce product descriptions, healthcare reporting, surveillance systems, and intelligent media management.

The project also highlights the growing importance of Vision-Language Models in modern Artificial Intelligence research. By integrating visual understanding with language generation, the system lays the foundation for more advanced multimodal applications such as visual question answering, video narration, human-computer interaction, and embodied AI systems.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” International

Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.