International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Audio Enabled Object Detection System with NLP based Interaction

Aishwarya Pralhad Chakrapani[1], Aryan Mayur Dhuru[2], Shraawani Prakash Lattoo[3], Parth Samir Dalvi[4], Rohit Sanjay Shahane[5], Prof. Poonam Sunil Jadhavar[6]

[1,2,3,4,5] *Department of Computer Engineering, Padmabhooshan Vasantdada Patil Institute of Technology, Pune*

ABSTRACT - The project aims to develop an innovative object detection system that uses audio input and output to enhance user engagement and provide detailed product information. The system employs deep learning techniques for real-time object identification, generating an audio-based output with the object's name. In order to enable users to interact with the system verbally, it also incorporates Natural Language Processing (NLP), which interprets and processes speech in order to identify the object. The system also includes an online search module, which provides users with descriptions of the object's attributes and potential applications. The system undergoes rigorous testing and optimization to ensure accuracy and responsiveness.

The project employs user-friendly technology that can assist people with low vision or disabilities, providing a practical and time-saving alternative for seeking instant information about objects. In order to make object recognition and information retrieval a smooth and inclusive experience for all users, this audio-enabled system is a first step towards bridging the gap between artificial intelligence and human-technology interaction.

***Key Words*:** Object Detection, Natural Language Processing, YOLOv4, Google Gemini, MeaningCloud, Wikipedia library

1. INTRODUCTION

The proposed project aims to create a system that converts visual information into auditory feedback, enabling individuals to navigate their surroundings independently. By employing real-time object detection using image processing and machine learning, the system alerts users to obstacles through audio output. This technology aids in daily activities, social interaction, and accessing printed information. Additionally, the survey paper explores conversational AI's advancements and applications, including a project integrating object detection, natural language processing, and online product information retrieval. Traditional aids like canes offer some support, but technological advancements present an opportunity to revolutionize navigation and information access. The system combines hardware such as sensors and cameras with deep learning algorithms to analyze environmental data and provide auditory guidance. By prioritizing mobility and information access, the project aims to promote inclusivity and independence, leveraging AI to enhance individuals' lives.

2. RELATED WORK

Vision is a vital sense for living beings, but millions worldwide grapple with visual impairment, posing challenges in navigation, information access, and communication. The lives of visually impaired individuals are often fraught with challenges, particularly in independently identifying and navigating their surroundings. The utilization of computer vision-based object detection techniques has emerged as a promising avenue for assisting the visually impaired by swiftly detecting and categorizing objects in real time [3]. The aim of the proposed project is to transform the visual landscape into an auditory one, alerting blind individuals to objects in their vicinity; improve accessibility and usability for people with visual impairment through advances in computer vision and machine learning technologies. Through a real-time object detection system, visually impaired individuals can navigate independently, with image processing and machine learning identifying objects via a camera and conveying their location through audio descriptions [1].

We tackle the challenge of integrating vision and language to comprehend objects in greater detail. Our approach hinges on utilizing object descriptions to enhance understanding. The utilization of computer vision-based object detection techniques has emerged as a promising avenue for assisting the visually impaired by swiftly detecting and categorizing objects in real-time [3]. This study introduces a pioneering real-time object detection and audio feedback system tailored to empower visually impaired individuals in recognizing and navigating their environment autonomously We introduce two novel architectures aimed at addressing two interrelated tasks: object captioning and natural language-based object retrieval. Object captioning involves detecting an object and generating its description simultaneously, while object retrieval entails localizing an object based on a given query [6]. The utilization of computer vision-based object detection techniques has emerged as a promising avenue for assisting the visually impaired by swiftly detecting and categorizing objects in real-time. This study introduces a pioneering real-time object detection and audio feedback system tailored to empower visually

impaired individuals in recognizing and navigating their environment autonomously.

Our method harnesses the YOLO_v4 algorithm in tandem with the MS COCO dataset to rapidly detect and classify objects, subsequently providing corresponding audio feedback [3]. A neural network consists of input with minimum one hidden and output layer. Multiple object dataset which consists of classes of images such as Car, truck, person, and two-wheeler captured during RGB and grayscale images. The dataset is composed (image and video) of varying illumination. YOLO model variants such as YOLOv3 is implemented for image and YOLOv4 for video dataset. Obtained results show that the algorithm effectively detects the objects approximately with an accuracy of 98% for image dataset and 99% for video dataset [14].

The transformative potential of a simple Object Detection model, utilizing Python libraries and APIs, to aid visually impaired individuals in overcoming these challenges. The proposed methodology introduces a Machine Learning model designed to detect object locations surrounding the individual and provide voice feedback Training the model utilizes the MS-COCO dataset to improve its effectiveness in recognizing diverse objects with various aspects [8]. Despite numerous existing solutions, many are either costly or ineffective. This paper introduces a groundbreaking system designed to assist visually impaired individuals in object localization. Utilizing image-to-speech technology through smartphone cameras, the system offers voice-guided directions. A pivotal innovation is the use of the user's hand as a reference point. By detecting the user's hand and identifying target objects within the camera's view, the system guides users to the object's location using image-to-speech techniques [7]. Furthermore, we have devised algorithms to deliver directional guidance and distance information, complemented by auditory feedback.

3. DATASET

For a project aimed at aiding visually impaired individuals in navigating their environments, training data should encompass various data types related to the system's intended tasks. This includes a diverse set of images and videos capturing different environments and scenarios, such as indoor and outdoor scenes, varying lighting conditions, and different types of objects, obstacles, and landmarks. Additionally, annotated object datasets with bounding boxes or segmentation masks for relevant navigation objects like pedestrians, vehicles, obstacles, traffic signs, and navigation aids (e.g., stairs, ramps, elevators) are essential. Audio data, including recordings of environmental sounds and cues such as traffic noises, pedestrian crossings, public announcements, and verbal instructions, also play a crucial role in training the system.

Leveraging pre-trained models and transfer learning on large-scale datasets can further enhance the training process by fine-tuning or using feature extractors tailored to the project's specific tasks. Integrating datasets like the COCO (Common Objects in Context) dataset, which offers a comprehensive collection of annotated images, can significantly improve object detection and recognition models. By combining COCO data with other relevant datasets, the project can enrich its training process, enabling the system to better recognize and interpret objects in various contexts. This comprehensive approach ensures that the algorithms and models are well-equipped to handle the complexities of real-world environments encountered by visually impaired individuals, thereby enhancing the system's usability and efficacy.

4. METHODOLOGIES

4.1. System Architecture

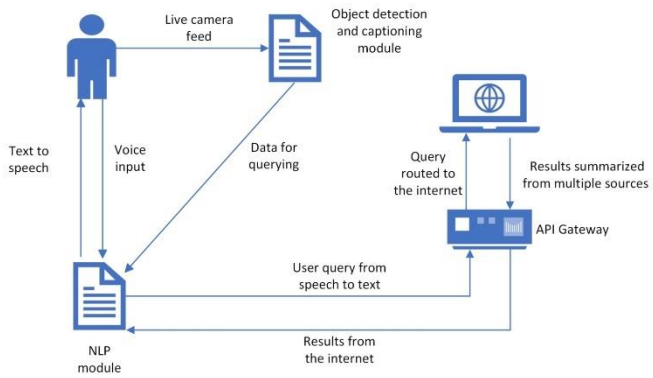


Fig -1: System Architecture

The system integrates a computer vision module and a natural language processing (NLP) module to provide a comprehensive query response. The computer vision module utilizes the YOLOv4 object detection model to analyse the live video feed from the camera. This model identifies and localizes objects within the frame, and generates textual captions or descriptions of the detected objects. The NLP module is responsible for processing the user's voice input. It leverages the Google Gemini API and Wikipedia library in python to generate a detailed response to the user's query. To concisely summarize this lengthy response, the system then employs the MeaningCloud API. This API condenses the Gemini-generated text into a succinct 2-line summary, which is the final output presented to the user.

This integrated approach combines computer vision techniques and natural language processing capabilities

to deliver a multimodal solution. The live camera feed provides visual information, which is annotated through object detection and captioning. The user's voice input is then processed by the NLP module, and the query response is formulated by aggregating data from both the visual and textual domains. The result is a concise, informative summary that leverages the strengths of multiple AI-powered technologies.

4.2. YOLOv4 Architecture

The block diagram of YOLOv4 is shown in Figure 2. These components are detailed as follows:

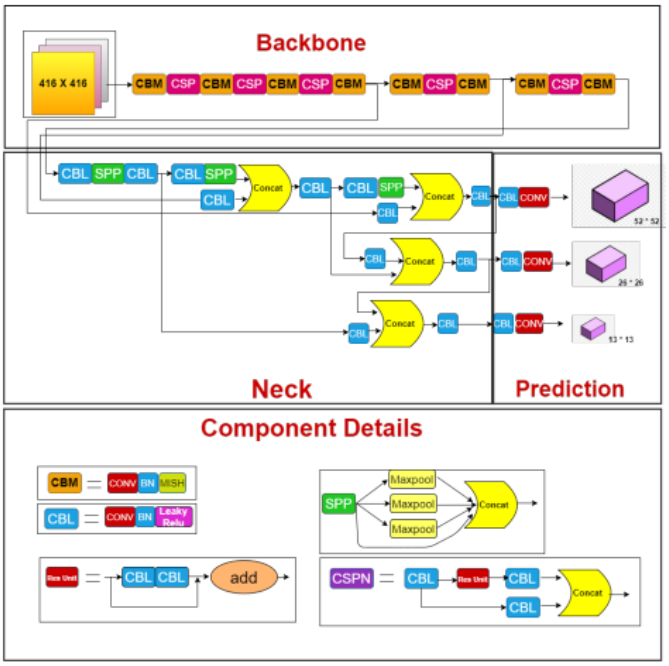


Fig -2: YOLOv4 Architecture

The backbone of YOLOv4 is CSPDarkNet53 which is a combination of YOLOv3 Darknet53 and Cross Stage Partial Network (CSPNet) in 2019. The addition of CSP blocks in Darknet-53 backbone increases the learning ability of CNN while keeping the memory consumption and calculations lower. The backbone of YOLOv4 uses Mish activation function that has shown good performance as compared to other activation functions. Mathematically, Mish activation function can be described as follows:

$$f(x) = x * tanh\ (ln\ (1 + exp\ x)) \quad (1)$$

The backbone consists of five CSP modules. Each CSP module reduces the image size by half. Hence, if input image has dimension 608x608, then the feature map obtained after backbone will be 416/32 = 13x13.

The neck of YOLOv4 consists of spatial pyramid pooling (SPP) and CBL blocks. The SPP module uses the maximum pooling method of k = [1x1, 5x5, 9x9, 13x13] and then concatenates the feature maps of different scales as shown in Figure. 2. The neck of YOLOv4 uses feature pyramid network (FPN) + path aggregation network (PAN) structure. This structure communicates semantic features from top to bottom to obtain the prediction feature map. The prediction step consists of YOLO layers to obtain the feature maps of three different scales i.e 13x13, 26x26 and 52x52. Predicting feature maps at different scales helps to recognize objects of small, medium and large sizes in the image.

4.3. Google Gemini Architecture

Gemini 1.5 Pro is a sparse mixture-of-expert (MoE) Transformer-based model that builds on Gemini 1.0's research advances and multimodal capabilities. Gemini 1.5 models were designed for high efficiency and low latency at all context lengths. For most queries, the time per output character dominates total request latency during transformer inference and this compounds for applications using chained prompts and agentic workflows. Gemini 1.5 models are trained on multiple 4096-chip pods of Google's TPUv4 accelerators, distributed across multiple datacenters, and on a variety of multimodal and multilingual data[16].

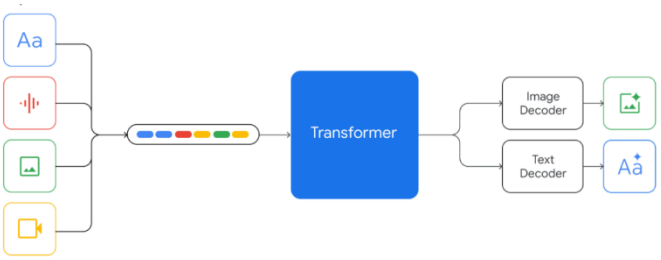


Fig -3: Google Gemini Architecture

4.4. MeaningCloud

The MeaningCloud Summarization API uses extractive summarization, which involves evaluating incoming text to select and provide the most informative sentences. The method starts by separating the input text into distinct sentences. Each sentence is then assessed using a combination of algorithms, including TextTeaser and TextRank, which take into account criteria such as the sentence's position in the text, the existence of emphasized words (italics or bold), numbers, and specific keywords or phrases. The summary is created by selecting the sentences with the best scores, with minimal post-processing tweaks for capitalization and punctuation.

The API is accessible through the endpoint 'POST https://api.meaningcloud.com/summarization-1.0' and communicates via HTTPS on port 443, ensuring secure data transmission. The API's response is in JSON format, resulting in a structured and easy-to-parse output that

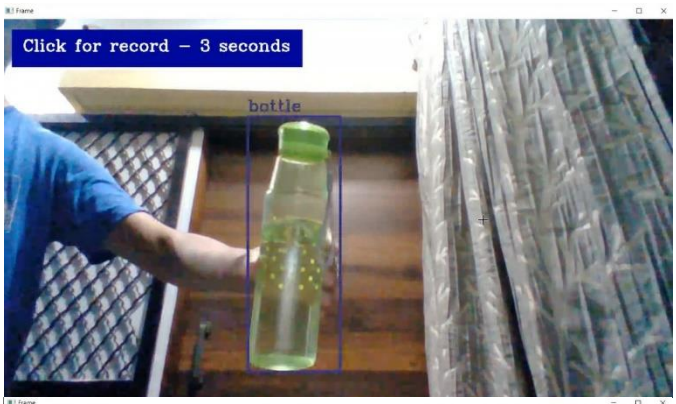
contains the created summary and request status information.

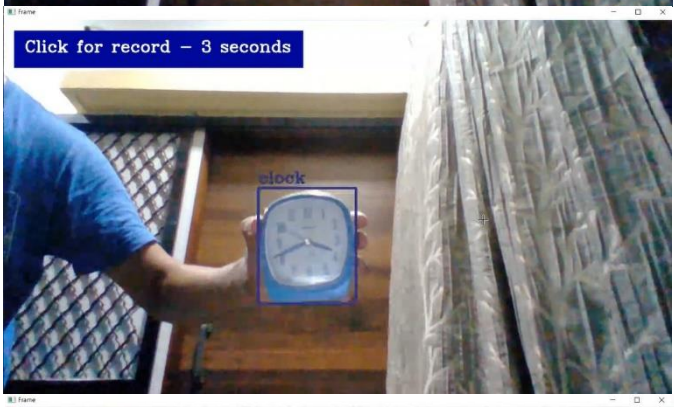
4.4. Wikipedia Library

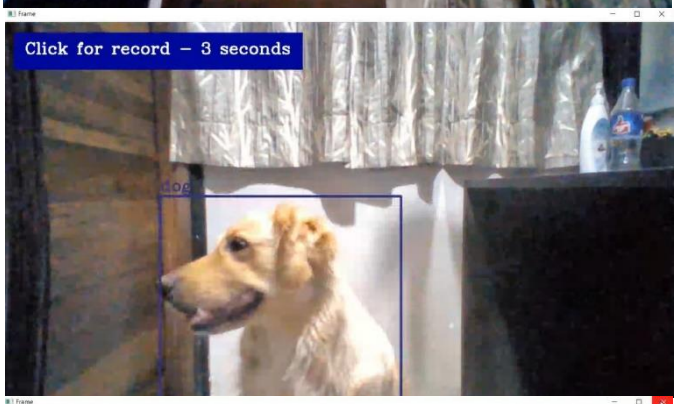
The Python Wikipedia library wraps the Wikipedia API, allowing Python applications to access Wikipedia material programmatically. It talks with Wikipedia servers using the HTTP protocol, issuing and receiving HTTP requests and answers. This allows for seamless interaction with Wikipedia's large library of information. The library offers developers useful functions and methods for retrieving Wikipedia articles, summaries, search results, and other resources, making tasks like data retrieval and analysis easier. Its use of Wikipedia's API simplifies the process of obtaining and editing Wikipedia material, providing developers with a strong tool for incorporating Wikipedia data into their applications. However, like any external API requirement, its reliance on network connectivity raises issues such as network latency and possibly API usage constraints. Despite these limitations, the Python Wikipedia library is yet a useful resource for developers looking to incorporate Wikipedia data into their projects.

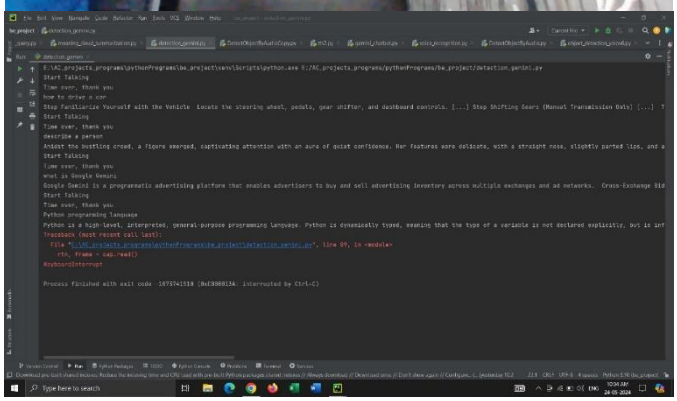
5. RESULTS

- Conducted real-time tests in controlled environments (e.g., homes, offices) and uncontrolled environments (e.g., streets, parks).
- The system successfully detected and identified objects with a latency of less than 5 second.
- Users reported that the auditory feedback was clear and helpful with object identification.
- The audio-based object detection system demonstrated robust performance, significantly helping blind individuals by enhancing their environmental awareness and independence.
- Real-time testing confirmed its intended result in both controlled and uncontrolled environments, with high user satisfaction.
- The audio-based object detection project has yielded substantial positive outcomes across various metrics, environments, and user demographics.









PARAMETER	GEMINI	WIKIPEDIA
Model	Gemini object detection systems are often based on advanced deep learning models such as Faster R-CNN, YOLO (You Only Look Once), or SSD (Single Shot MultiBox Detector).	Wikipedia object detection may refer to basic, general-purpose models available through open-source platforms or API services like Google Cloud Vision or AWS Rekognition.
Accuracy	These systems are designed to achieve high accuracy in object detection tasks by leveraging large-scale annotated datasets and sophisticated training techniques.	These systems are designed for general use and may not be as finely tuned or accurate as specialized systems like Gemini.
Speed	While some Gemini models like YOLO prioritize speed, others like Faster R-CNN emphasize precision, often trading off speed for accuracy.	Typically optimized for general usage, providing a balance between speed and accuracy suitable for broad applications.
Applications	**Industry Use**: Gemini object detection is widely used in various industries, including autonomous driving, surveillance, healthcare, and retail.	**Integration**: It can be integrated into web services and applications where moderate accuracy is sufficient and ease of use is a priority.
Training Data and Customization	These systems are typically trained on large, diverse datasets like COCO (Common Objects in Context) or custom datasets tailored to specific use cases.	These models are usually trained on large, publicly available datasets, providing broad but not necessarily deep knowledge of specific object categories.
Customization	Gemini systems can be fine-tuned on domain-specific data to improve performance on particular tasks or environments.	Limited customization options compared to specialized systems. Users often rely on the pre-trained capabilities of the models.
Technology	They often use machine learning frameworks such as TensorFlow, PyTorch, or Keras for model development and training.	Often built using accessible machine learning libraries and services that prioritize ease of integration over cutting-edge performance.
Deployment	These models can be deployed on various platforms, from cloud servers to edge devices, depending on the application requirements.	Usually available through APIs and cloud services, making them easy to use but potentially limited in customization and control.

Table -1: System Comparison

5. FUTURE WORK

To enhance the capabilities of our real-time object detection system for visually impaired individuals, several advancements can be considered. First, increasing the variety of detectable objects and environmental sounds, such as specific types of furniture, electronic devices, and natural sounds like water sources and birds, will broaden the system's utility. Advanced noise filtering techniques should be developed and integrated to improve performance in noisy environments, such as busy streets, markets, and public transport.

Employing edge computing to process data locally on the device can reduce latency and enhance real-time detection. Integrating the system with smart glasses equipped with built-in microphones and speakers can provide a seamless user experience. Additionally, forming partnerships with associations that support blind and visually impaired individuals can help gather valuable feedback, conduct user trials, and promote the adoption of the technology, ensuring it meets the users' needs effectively.

6. CONCLUSION

In summary, the creation of the "Audio-Enabled Object Detection System with NLP-based Interaction" signifies a significant advancement in merging artificial intelligence with human-technology interaction, particularly in enhancing accessibility and usability for individuals with visual impairments or limited visual abilities. This initiative strives to create a comprehensive solution empowering users to independently gather crucial information, interact with their environment, and make informed decisions.

The system offers a seamless and user-friendly interface by amalgamating state-of-the-art technologies such as object detection, natural language processing (NLP), text-to-speech synthesis, and integration with advanced language models. Through voice-activated object inquiries, precise audio responses, and detailed product descriptions sourced from online platforms, users can better understand their surroundings and gain access to commercial, educational, and informational resources.

Looking ahead, this system holds potential applications in assistive technology, education, travel, home automation, and beyond. It lays the foundation for future

exploration, innovation, and societal impact as technology progresses. By enabling individuals with visual impairments to engage more meaningfully with their environment, this project underscores the transformative potential of AI in fostering inclusivity and equity in society.

7. REFERENCES

1. Mahendru, M. and Dubey, S.K. (2021) 'Real time object detection with audio feedback using Yolo vs. Yolo_v3', *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* [doi:10.1109/confluence51648.2021.9377064.
2. Vaidya, S. *et al.* (2020) 'Real-time object detection for visually challenged people', *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. doi:10.1109/iciccs48265.2020.9121085.
3. Nguyen, A. *et al.* (2019) 'Object captioning and retrieval with natural language', *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. doi:10.1109/iccvw.2019.00316.
4. Jambhulkar, A.R. *et al.* (2023) 'Real-time object detection and audio feedback for the visually impaired', *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)* doi:10.1109/asiancon58793.2023.10269899.
5. Chaturvedi, R.K. *et al.* (2023) 'Visual object detection using audio data', *Journal of Physics: Conference Series*, 2664(1), p. 012006. doi:10.1088/1742-6596/2664/1/012006.
6. Li, J. *et al.* (2022) 'Lightweight real-time object detection system based on embedded AI Development Kit', *2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*. doi:10.1109/mlise57402.2022.00010.
7. Gupta, S. *et al.* (2022) 'Object detection with audio comments using Yolo V3', *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. doi:10.1109/icaaic53929.2022.9792755.
8. Denizgez, T.M. *et al.* (2021) 'Guiding visually impaired people to find an object by using image to speech over the smart phone cameras', *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. doi:10.1109/inista52262.2021.9548122.
9. Hitaish, K. *et al.* (2023) 'Voice assisted object detection for visually impaired', *2023 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. doi:10.1109/conecct57959.2023.10234781.
10. Islam, R.B. *et al.* (2023) 'Deep learning based object detection and surrounding environment description for visually impaired people', *Heliyon*, 9(6). doi:10.1016/j.heliyon.2023.e16924.
11. Calvin, R. and Suresh, S. (2021) 'Image captioning using convolutional neural networks and recurrent neural network', *2021 6th International Conference for Convergence in Technology (I2CT)*. doi:10.1109/i2ct51068.2021.9418001.
12. Wang, C.-Y., Bochkovskiy, A. and Liao, H.-Y.M. (2023) 'Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors', *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr52729.2023.00721.
13. Kim, J.U. and Man Ro, Y. (2019) 'Attentive layer separation for object classification and object localization in object detection', *2019 IEEE International Conference on Image Processing (ICIP)*. doi:10.1109/icip.2019.8803439.
14. Kumar B., C., Punitha, R. and Mohana (2020) 'Yolov3 and Yolov4: Multiple object detection for surveillance applications', *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. doi:10.1109/icssit48917.2020.9214094.
15. Ali, S. *et al.* (2021) 'Improved Yolov4 for aerial object detection', *2021 29th Signal Processing and Communications Applications Conference (SIU)*. doi:10.1109/siu53274.2021.9478027.
16. Google (2024) Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context