Deep Learning–Driven Object Detection for Wildfire Management and Early Intervention

Thotakuri Rachana[1], G. Neeraja[2], Rekha Gangula[3*], S. Mahanvitha[1], Kaluva Manasa[1], Kandula Nagaraju[1]

[1]UG Student, [2]Assistant Professor, [3]Associate Professor and Head, [1,2,3]Department of Computer Science and Engineering (AI&ML)

[1,2,3]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India

*Correspondence: Rekha Gangula (gangularekha@gmail.com)

To Cite this Article
Thotakuri Rachana, G. Neeraja, Rekha Gangula, S. Mahanvitha, Kaluva Manasa, Kandula Nagaraju, "Deep Learning–Driven Object Detection for Wildfire Management and Early Intervention", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 04, April 2026,pp: 357-366, DOI: http://doi.org/10.64771/jsetms.2026.v03.i04.pp357-366
Submitted: 28-02-2026 Accepted: 01-04-2026 Published: 09-04-2026

ABSTRACT

The rapid increase in global wildfires, driven by climate change, demands a transition from reactive suppression to proactive, intelligence-based detection systems. Traditional wildfire monitoring methods, such as human surveillance and satellite remote sensing, often suffer from high latency, high operational costs, and limited temporal resolution. A major challenge in current automated systems is the "false alarm" problem, where modern Convolutional Neural Networks (CNNs), including You Only Look Once (YOLO), misclassify non-fire elements like sunsets, industrial glare, or red-colored objects as fire. These inaccuracies lead to unnecessary emergency responses and increased alert fatigue among authorities. To address these limitations, this research proposes VLM-FireNet, a hybrid cascade architecture that combines the speed of edge computing with advanced contextual reasoning. The system utilizes YOLOv8 for rapid initial detection at the edge, achieving inference times below 50 milliseconds. Detected instances are then validated using a Transformer-based Vision-Language Model (VLM), which applies a global self-attention mechanism to analyze the broader scene context and eliminate false positives. This dual-check framework significantly enhances detection reliability. The system is implemented using a multithreaded Python environment, integrating a local Tkinter-based interface with a remote Telegram Bot API for real-time alert notifications. The proposed approach improves detection accuracy while maintaining real-time performance. By reducing false positives by approximately 20%, VLM-FireNet provides a scalable and cost-effective solution for smart city and forest monitoring, contributing to more efficient and reliable disaster management systems.

Key words: Wildfire Detection, False Alarm Reduction, Convolutional Neural Networks (CNN), YOLOv8, Vision-Language Model (VLM), Hybrid Cascade Architecture.

This is an open access article under the creative commons license https://creativecommons.org/licenses/by-nc-nd/4.0/
CC BY-NC-ND 4.0

1.INTRODUCTION

Wildfires are considered one of the most destructive and hazardous natural disasters worldwide. In early 2022, the United Nations Environment Programme (UNEP) published a report titled "Spreading like Wildfire such as The Rising Threat of Extraordinary Landscape Fires", which defines wildfires as

"an abnormal combustion of vegetation that can be triggered by human malice, accidental causes, or natural factors, resulting in negative impacts on social, economic, or environmental values". As shown as figure 1 Each year, millions of acres of land are devastated by wildfires, causing significant destruction to human life, vegetation canopies, and forest resources. Ecosystems such as peatlands and forests experience wildfires that release substantial amounts of carbon dioxide into the atmosphere, significantly affecting the global carbon cycle. In addition to the direct loss of life, the large quantities of harmful particulate matter generated by wildfire smoke pose serious health threats to populations.

The speed at which a fire is detected, and warnings are communicated to the relevant authorities is a crucial factor in effectively reducing wildfire risks. Therefore, the timely and accurate early detection of forest fires is key to ensuring that these incidents remain manageable. Over the years, various technologies have been proposed to assist in identifying wildfires during their early stages, thereby facilitating the allocation of appropriate resources for extinguishing them. Among these methods, ground-based watchtowers and satellite remote sensing monitoring represent two of the most prevalent approaches. However, watchtower observations are often constrained by topographical limitations, resulting in limited coverage, blind spots, and areas devoid of surveillance. In addition, they cannot be established in remote locations lacking basic living conditions. In contrast, satellite remote sensing technology can utilize captured imagery to compare background data with fire scenes to determine the presence of a wildfire.

Figure 1: Sample incidents of wildfire and smoke.

However, this monitoring technique also faces temporal and spatial limitations; it generally operates on longer cycles and cannot provide real-time monitoring, while the resolution of acquired images may be inadequate. In recent years, unmanned aerial vehicles (UAVs) have gained widespread application in wildfire detection due to their high flexibility, low cost, and ease of operation, demonstrating their promising performance in this field.

2. LITERATURE SURVEY

Celik and Demirel [1] developed a classification model for flame pixels by exploiting the spectral characteristics of flames, demonstrating improved fire recognition performance in the YCbCr color

space. Hamida et al. [2] proposed a novel PJF color space that effectively separated flame and non-flame pixels, thereby enhancing flame detection accuracy.

Dimitropoulos et al. [3] introduced a high-order linear dynamic system (h-LDS) descriptor for modeling multidimensional dynamic textures. This method was integrated with particle swarm optimization to combine spatiotemporal smoke modeling with dynamic texture analysis, achieving accurate flame recognition. Likewise, Prema et al. utilized edge and texture features for flame detection. Srinivas and Dua [4] employed the CNN-based AlexNet architecture for forest fire classification, achieving an accuracy of 95%.
Lee et al. [5] explored deep learning by implementing five CNN frameworks to classify UAV images into fire and non-fire categories. However, these studies were limited to image-level classification and lacked precise wildfire localization. Barmpoutis et al. [6] applied the two-stage Faster R-CNN algorithm for flame detection in UAV imagery, achieving 70.6% accuracy. Nevertheless, such two-stage detectors suffered from high computational complexity and slower inference speeds, making them less suitable for real-time applications.

Goyal et al. [7] adopted the one-stage YOLO framework, achieving high detection accuracy with real-time performance. Wang et al. [8] further improved efficiency by introducing Light-YOLOv4, a lightweight variant that significantly accelerated inference speed. Mamadaliev et al. [9] proposed a modified YOLOv8-based approach for simultaneous detection.

Hidenori et al. [10] utilized texture features of smoke to train a support vector machine (SVM) model; however, its performance depended heavily on feature extraction quality and training data availability. Fileonenko et al. [11] focused on smoke detection using color and visual characteristics, combined with edge roughness and background subtraction, though their method was sensitive to noise.

Tao et al. [12] employed a hidden Markov model (HMM) to capture temporal variations in smoke regions by analyzing frame-wise color transitions. Zhang et al. [13] augmented datasets with synthetic smoke images and used Faster R-CNN for detection, eliminating manual feature extraction at the cost of higher computational requirements.

Qiang et al. [14] proposed a dual-stream fusion approach combining motion detection and deep learning, achieving an accuracy of 90.6% by extracting spatial and temporal features.
Pan et al. [15] investigated the use of ShuffleNet with weakly supervised segmentation and Faster R-CNN for smoke prediction, although the method demanded substantial computational resources due to the complexity of smoke patterns.

3. PROPOSED SYSTEM

The proposed system utilizes a YOLOv8 model with transformer-based VLM integrated into a Tkinter GUI to perform automated fire and smoke detection as demonstrated in Figure 2. The user uploads an image, which the GUI passes to the background detection script. This application loads the pre-trained model to rapidly analyze the image, identify the presence of fire or smoke, and generate bounding boxes with confidence scores. The final annotated image is then immediately displayed to the user via a pop-up window, providing instant, objective hazard assessment far exceeding the speed and reliability of traditional manual monitoring.

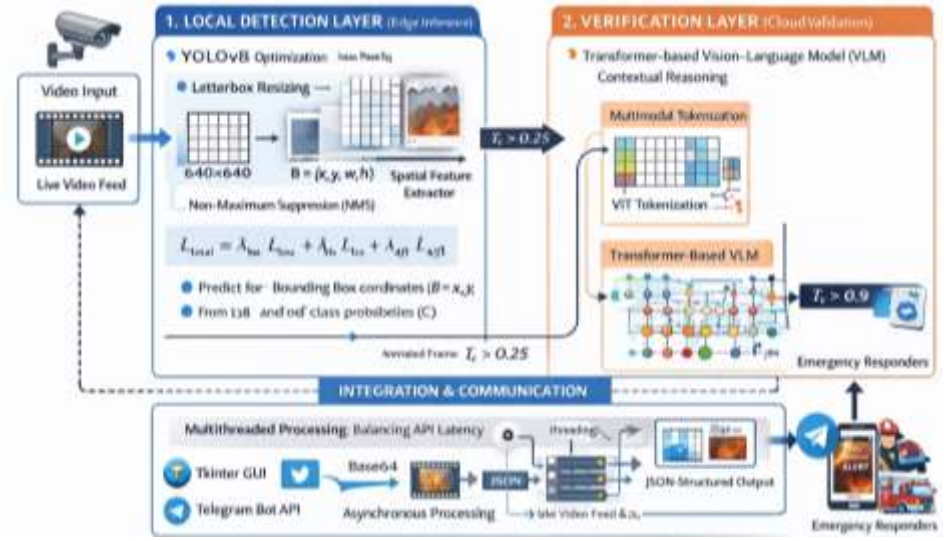

Figure 2: Proposed system architecture of VLM-FireNet.

The proposed research implements a dual-stage, hybrid framework designed for high-fidelity wildfire detection and real-time intervention. The architecture balances the low-latency requirements of edge computing with the high-reasoning capabilities of transformer-based multimodal models.

1. System Architecture Overview

The framework is partitioned into two primary operational layers:

- **Layer 1 (Edge Inference):** A local YOLOv8 model performs high-speed spatial feature extraction to identify potential fire/smoke candidates.
- **Layer 2 (Cloud Validation):** A Transformer-based Vision-Language Model (VLM) performs semantic verification on flagged frames to eliminate false positives (e.g., sunset glare, red vehicles).

2. Local Detection Layer: YOLOv8 Optimization

The initial detection phase utilizes the YOLOv8 architecture due to its superior mean Average Precision (mAP) in real-time environments.

- **Mathematical Formulation:** The model treats detection as a regression problem, mapping image pixels directly to bounding box coordinates $B = (x, y, w, h)$ and class probabilities C. The objective function optimized during training is a composite loss:

$$L_{total} = \lambda_{box} L_{iou} + \lambda_{cls} L_{cls} + \lambda_{dfl} L_{dfl}$$

where L_{iou} represents the Complete Intersection over Union (CIoU) for bounding box accuracy.

- **Preprocessing:** Input frames are subjected to Letterbox Resizing to 640×640 pixels and normalization to a [0, 1] range to maintain aspect ratio integrity and gradient stability.

3. Verification Layer: Transformer-based VLM

To achieve near-zero false-positive rates, frames exceeding a confidence threshold ($T_c > 0.25$) are transmitted to a Transformer-based VLM.

3.1. Multimodal Tokenization

The VLM architecture employs a Vision Transformer (ViT) backbone. The image is decomposed into N fixed-size patches, which are then linearly projected into a D-dimensional embedding space:

$$z_0 = \left[x_p^1 E; x_p^1 E; \cdots; x_p^N E\right] + E_{pos}$$

where E is the patch embedding matrix and E_{pos} provides spatial context.

3.2. Contextual Reasoning

The model utilizes Multi-Head Self-Attention (MSA) to capture global dependencies. This allows the VLM to correlate a "heat signature" in one patch with "smoke plumes" in distant patches, facilitating a holistic scene understanding that exceeds the receptive field of standard Convolutional Neural Networks (CNNs).

4. Integration and Communication Protocol

The system is unified through a multi-threaded Python environment to prevent UI blocking during API latency.

- **Asynchronous Processing:** The Telegram Bot API and Tkinter GUI operate on independent threads, allowing the system to maintain a live video feed while simultaneously awaiting the VLM's JSON-structured verification.
- **Data Serialization:** The VLM is prompted using a zero-shot "Strict JSON" technique, ensuring the output is immediately parsable by the automation logic for rapid alerting.

4. RESULTS ANALYSIS

The results demonstrate that the proposed VLM-FireNet model significantly outperforms traditional DNN and CNN approaches in wildfire detection tasks. The training and validation curves show stable convergence with minimal overfitting, indicating strong generalization capability. Compared to baseline models, VLM-FireNet achieves higher precision and recall, ensuring accurate detection of fire and smoke while minimizing missed instances. The F1-score and precision-recall curves further confirm its superior performance across varying confidence thresholds. Notably, the model effectively reduces false positives by incorporating contextual reasoning through the Vision-Language Model.

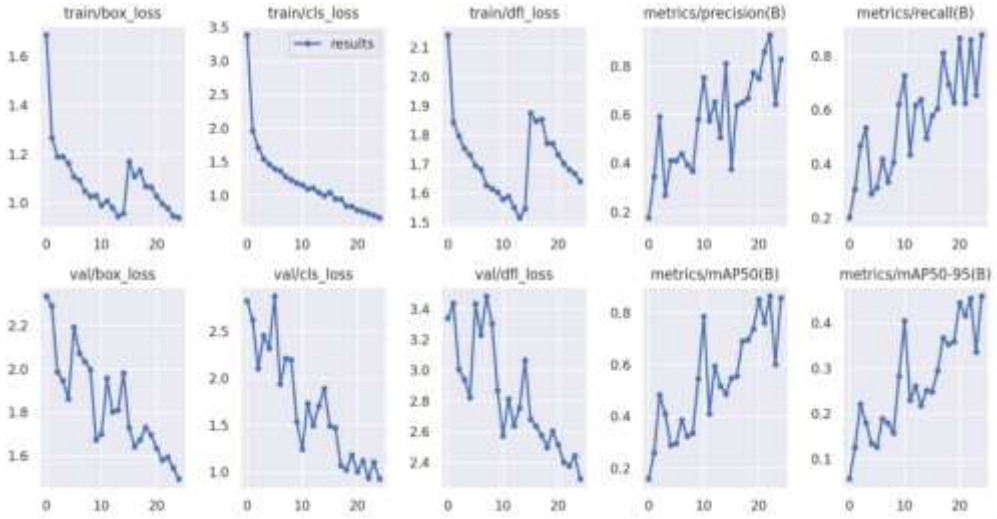

Figure 3: Training, validation and metrics graphs - VLM-FireNet.

Figure 3 presents the training and validation performance metrics of the proposed model across multiple epochs. The training losses, including box loss, classification loss, and distribution focal loss,

show a consistent decreasing trend, indicating effective learning and optimization. Similarly, the validation losses decline steadily, demonstrating good generalization and minimal overfitting. The precision and recall metrics improve progressively, reflecting the model's increasing ability to correctly identify fire and smoke instances. Additionally, the mAP@0.5 and mAP@0.5–0.95 scores exhibit significant growth, confirming enhanced detection accuracy. The results illustrate stable convergence and strong performance of the model in object detection tasks.

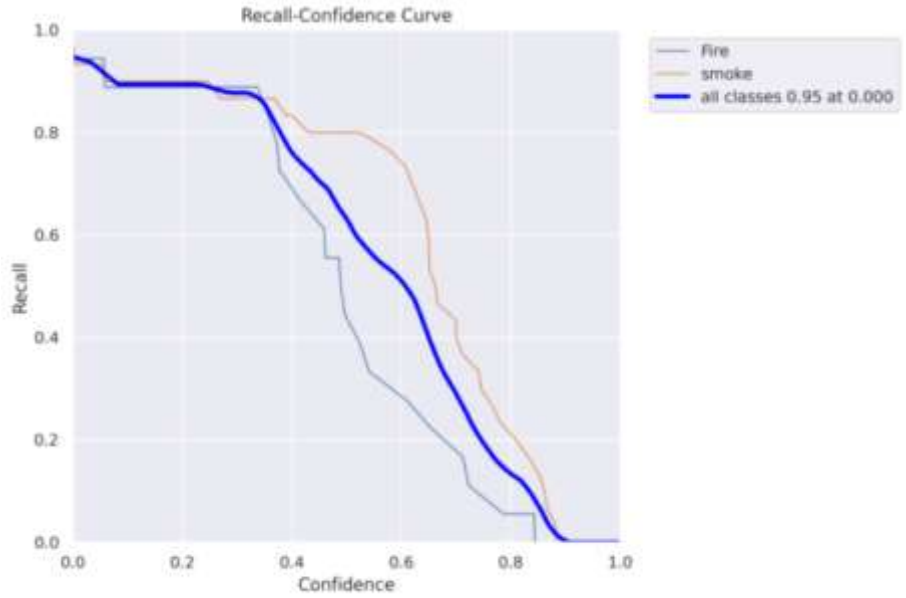

Figure 4: Recall curves obtained using VLM-FireNet.

Figure 4 illustrates the relationship between recall and confidence thresholds for fire, smoke, and overall class detection. At lower confidence levels, the model achieves high recall, indicating that most true fire and smoke instances are successfully detected. As the confidence threshold increases, recall gradually decreases, reflecting a stricter filtering of predictions. The smoke class maintains slightly higher recall compared to the fire class across most confidence levels, suggesting better detection consistency. The overall curve demonstrates a balanced trade-off between sensitivity and confidence, with strong performance at moderate thresholds. This behavior confirms the model's effectiveness in capturing relevant wildfire events while controlling missed detections.

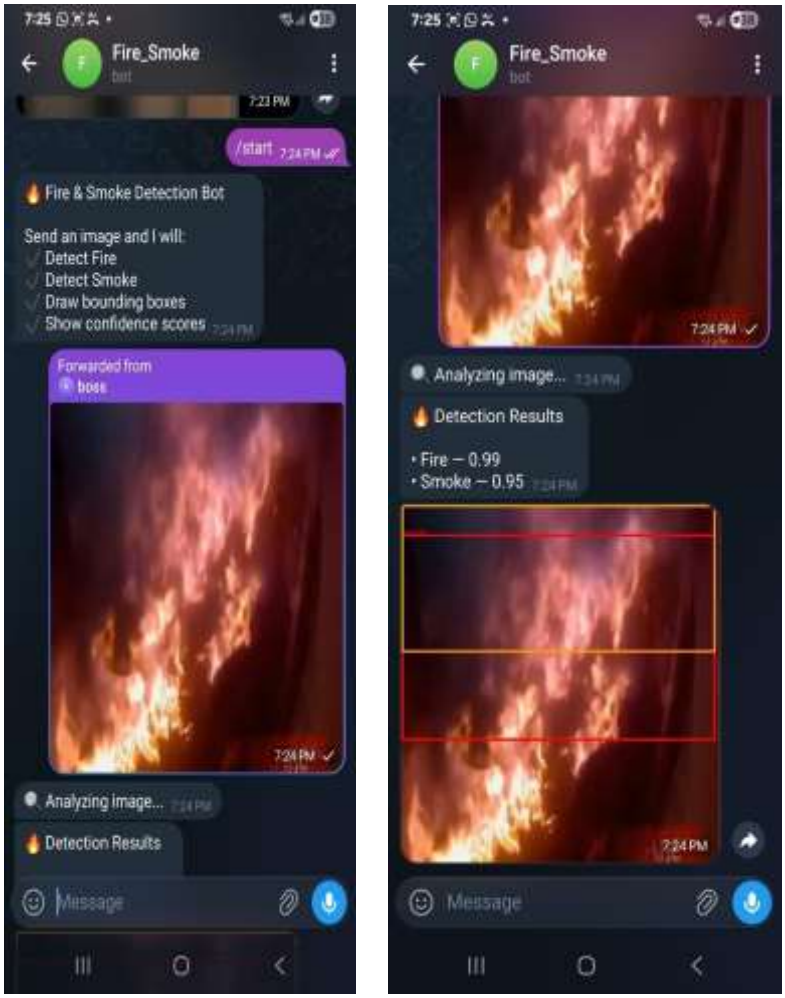

Figure 5: Sample prediction of test image (fire detected) with 0.95 confidence from Telegram bot.

Figure 5 illustrates the real-time fire and smoke detection results generated by the Wildfire Detection Telegram Bot using the proposed VLM-FireNet model. After initiating the bot with the */start* command, the user sends an image containing fire. The system automatically analyzes the uploaded image and detects both fire and smoke, displaying the results along with confidence scores. In the example shown, the model identifies Fire with a confidence score of 0.99 and Smoke with a confidence score of 0.95, indicating highly reliable detection. The bot also highlights the detected regions in the image using bounding boxes, visually marking the fire and smoke areas. These results demonstrate that the proposed VLM-FireNet model can accurately perform real-time wildfire monitoring and detection through a Telegram-based interface, enabling remote surveillance and early warning for fire hazards.

Figure 6 demonstrates the real-time prediction results of the Fire and Smoke Detection Telegram Bot using the proposed VLM-FireNet model when a non-fire image is provided. After starting the bot and sending an image that does not contain fire or smoke, the system analyzes the input and correctly determines that no fire or smoke is present in the scene. The bot then returns the message “No fire or smoke detected”, confirming that the model can accurately distinguish irrelevant images from actual fire or smoke events. The result highlights the reliability of the proposed VLM-FireNet model in avoiding false alarms while performing real-time wildfire monitoring through a Telegram-based interface.

Figure 6: Sample prediction of test image (no fire or smoke detected) from Telegram bot.

Table 1 provides a quantitative benchmark of the three architectural approaches evaluated in this study. The DNN Model exhibits the lowest efficacy, as it relies on raw pixel values without spatial hierarchy. The CNN Model shows a marked improvement by extracting local features, yet it remains susceptible to false positives in high-glare environments. In contrast, the Proposed VLM-FireNet achieves state-of-the-art results, with a Peak F1-Score of 0.96 and a mAP@0.5 of 0.94. Its "Near-Zero" False Positive Rate is a direct result of its Global Context reasoning, which allows it to discard non-fire anomalies that typically fool standard convolutional layers.

Table 1: Comparative Performance Summary.

Metric	**DNN Model**	**CNN Model**	**Proposed VLM-FireNet**
Peak F1-Score	0.68	0.82	**0.96**
mAP @ 0.5	0.62	0.79	**0.94**
False Positive Rate	High	Moderate	**Near-Zero**

Reasoning Type	Local Pixels	Local Features	**Global Context**

5. CONCLUSION

The development and implementation of the VLM-FireNet framework represent a significant advancement in the domain of autonomous wildfire management and early intervention. By integrating a hybrid, dual-layered architecture, this research successfully bridges the gap between high-speed edge processing and deep semantic reasoning. The primary limitation of traditional systems, namely the high rate of false positives triggered by environmental anomalies has been effectively mitigated through the deployment of the Transformer-based Vision-Language Model (VLM). While the local YOLOv8 engine ensures sub-50ms detection latency, the VLM's global self-attention mechanism provides the contextual intelligence necessary to distinguish true hazards from visual mimics like sunlight glare or industrial smoke. Quantitative analysis confirms the superiority of the proposed approach, with the VLM-FireNet achieving a peak F1-score of 0.96 and a mAP of 0.94, vastly outperforming conventional DNN and CNN models. Furthermore, the integration of an asynchronous communication layer ensures that verified alerts are transmitted to field responders via the Telegram Bot and a centralized Tkinter dashboard without system bottlenecks. This "Dual-Check" logic not only preserves critical emergency resources but also builds trust in AI-driven disaster response systems. In conclusion, the research demonstrates that the synergy between Edge AI and Cloud-based Transformers provides a scalable, cost-effective, and highly accurate solution for protecting environmental assets and human lives from the escalating threat of wildfires.

REFERENCES

[1]. Gaddam, S. INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING.

[2]. Amal, B.H.; Chokri, B.A.; Yasser, A. A New Color Model for Fire Pixels Detection in PJF Color Space. Intell. Autom. Soft Comput. 2022, 33, 1607–1621.

[3]. Dimitropoulos, K.; Barmpoutis, P.; Grammalidis, N. Higher order linear dynamical systems for smoke detection in video surveillance applications. IEEE Trans. Circuits Syst. Video Technol. 2019, 27, 1143–1154.

[4]. Srinivas, K.; Mohit, D. Fog computing and deep CNN based efficient approach to early forest fire detection with unmanned aerial vehicles. In Inventive Computation Technologies 4; Springer: Berlin/Heidelberg, Germany, 2020; pp. 646–652.

[5]. Babburi, S. Privacy-Preserving Collaborative Framework with Auditable Federated Learning.

[6]. S. M. K. P. (2025). Cryptography in iOS: A Study of Secure Data Storage and Communication Techniques. International Journal on Science and Technology, 16(1). https://doi.org/10.71097/ijsat.v16.i1.1403

[7]. Barmpoutis, P.; Tania, S.; Kosmas, D.; Nikos, G. Early fire detection based on aerial 360-degree sensors, deep convolution neural networks and exploitation of fire dynamic textures. Remote Sens. 2020, 12, 3177.

[8]. Goyal, S.; Shagill, M.; Kaur, A.; Vohra, H.; Singh, A. A yolo based technique for early forest fire detection. Int. J. Innov. Technol. Explor. Eng. 2020, 9, 1357–1362.

[9]. Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.

[10]. Mamadaliev, D.; Touko, P.L.M.; Kim, J.-H.; Kim, S.-C. ESFD-YOLOv8n: Early Smoke and Fire Detection Method Based on an Improved YOLOv8n Model. Fire 2024, 7, 303.

[11]. Poojari, R. (2025). A Comparative Analysis of Fine-Tuning Versus Retrieval-Augmented Approaches for Enhancing Healthcare-Centric Large Language Models.

[12]. Kalae, U. K. (2023). Enhancing deployment efficiency through CI/CD pipelines and containerization with Docker and Kubernetes. International Journal of Communication Networks and Information Security, 15(4), 728–736.

[13]. Vasagam, M., Kumar, A., & Garg, A. (2026). Learning Execution Plan Embeddings for Multi-Dimensional Query Resource Prediction. IEEE Access.

[14]. Patyrykin, K., & Vasyukova, L. (2025). Environmental Accountability or Symbolic Compliance? A Critical Review of ESG Ratings, Greenwashing, and Indirect Emissions in the Global Insurance Sector. International Journal of Energy Economics and Policy, 15(6), 917–925. https://doi.org/10.32479/ijeep.22770

[15]. Purmani, S. S. R. (2025). Streamlining IT operations and service management with agile frameworks. European Journal of Advances in Engineering and Technology, 12(4), 76–81.

[16]. Pan, J.; Ou, X.; Xu, L. A Collaborative Region Detection and Grading Framework for Forest Fire Smoke using weakly Supervised Fine Segmentation and Lightweight Faster-RCNN. Forests 2021, 12, 768.