Semantic-Aware Dual-Stage Visual Intelligence for Robust Wildfire Detection in Real-World Environments

J. Srikanth[1], Kadari Srihitha[2], K Varsha Sree[2], Maske James[2], Kavati Santhosh[2]

[1]Associate Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering

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

ABSTRACT

The growing severity and frequency of wildfires, largely driven by climate change, highlight the urgent need for intelligent and proactive monitoring systems. Traditional approaches, including manual surveillance and satellite-based observation, are constrained by delayed response times, high operational costs, and limited temporal resolution. A critical limitation in current automated detection systems is the high rate of false alarms. While modern Convolutional Neural Networks (CNNs) such as YOLO (You Only Look Once) enable fast, real-time detection, they often lack contextual awareness, leading to misclassification of visually similar elements like sunsets, artificial lighting, or coloured objects as fire. Such inaccuracies result in inefficient emergency responses and increased alert fatigue. To overcome these challenges, this work presents VLM-FireNet, a hybrid cascade framework that combines the speed of edge-based detection with the contextual reasoning capabilities of advanced multimodal models. In the proposed system, YOLOv8 is deployed at the edge to perform rapid initial detection with inference times below 50 milliseconds. These detections are subsequently verified using a Transformer-based Vision-Language Model (VLM), which leverages a global self-attention mechanism to analyse the broader scene context and filter out false positives effectively. The system is implemented within a multithreaded Python environment, integrating a Tkinter-based graphical interface with a Telegram Bot API for real-time remote alerting. The core contribution of this research is its dual-validation strategy, which enhances detection accuracy without compromising speed. Experimental evaluation shows that the proposed approach reduces false positives by approximately 20% while maintaining real-time performance. This hybrid methodology offers a scalable and efficient solution for AI-enabled IoT (AIoT) applications in wildfire detection and disaster management.

Keywords: Wildfire detection, edge computing, real-time monitoring, false alarm reduction, context-aware analysis, hybrid systems, AIoT, disaster management.

1. INTRODUCTION

Wildfires rank among the most destructive natural hazards, causing extensive damage to ecosystems, human life, and economic resources worldwide. In 2022, the United Nations Environment Programme (UNEP), through its report “Spreading Like Wildfire: The Rising Threat of Extraordinary Landscape Fires,” described wildfires as uncontrolled vegetation fires arising from natural causes, human negligence, or intentional actions, leading to severe environmental, social, and economic consequences. Every year, millions of acres of land are affected, resulting in the loss of biodiversity, destruction of forest cover, and long-term ecological imbalance, as shown in figure 1. Critical ecosystems such as forests and peatlands are especially vulnerable, as wildfires release significant amounts of carbon dioxide, intensifying global warming and disrupting the carbon cycle. Additionally, the emission of fine particulate matter from wildfire smoke poses serious health risks to nearby populations.

A key factor in mitigating wildfire damage is the speed and accuracy of detection, along with timely communication of alerts to emergency responders. Early detection plays a vital role in preventing small-scale fires from escalating into large, uncontrollable events. Over the years, various monitoring techniques have been developed to address this need. Traditional methods such as ground-based

watchtowers provide localized surveillance but are limited by terrain constraints, restricted visibility, and the inability to operate effectively in remote or inaccessible areas. Satellite-based remote sensing offers broader coverage by analysing large-scale imagery; however, it is constrained by lower temporal resolution, delayed data acquisition, and limited image clarity, making real-time detection challenging.

Figure 1: Sample incidents of wildfire and smoke.

To overcome these limitations, unmanned aerial vehicles (UAVs) have recently gained significant attention in wildfire monitoring. Due to their high mobility, cost-effectiveness, and ease of deployment, UAVs provide enhanced flexibility and improved observation capabilities. Their ability to capture high-resolution, real-time data makes them a promising solution for advancing early wildfire detection systems.

2. LITERATURE SURVEY

A wide range of techniques have been explored for wildfire detection, evolving from traditional image processing methods to advanced deep learning approaches. Early research primarily focused on color-based detection. For instance, Celik and Demirel [1] developed a flame pixel classification model based on spectral characteristics, demonstrating improved fire detection performance in the YCbCr color space. Similarly, Hamida et al. [2] introduced the PJF color space, which enhances the separation of flame and non-flame pixels, thereby improving detection accuracy.

Beyond color features, researchers have also leveraged texture-based methods. Dimitropoulos et al. [3] proposed a high-order linear dynamic system (h-LDS) descriptor to capture multidimensional dynamic texture features, integrating it with particle swarm optimization for effective flame recognition. Prema et al. further explored edge and texture-based features for identifying fire patterns. In parallel, the adoption of deep learning models began to gain traction. Srinivas and Dua [4] utilized the AlexNet CNN architecture for forest fire image classification, achieving an accuracy of 95%. Likewise, Lee et al. [5] evaluated multiple CNN frameworks to classify UAV images into fire and non-fire categories. However, these classification-based approaches are limited to image-level predictions and lack precise localization capabilities.

To address localization, object detection models have been introduced. Barmpoutis et al. [6] applied the two-stage Faster R-CNN algorithm for flame detection in UAV imagery, achieving an accuracy of 70.6%. Despite its effectiveness, the computational complexity of Faster R-CNN limits its real-time applicability. In contrast, Goyal et al. [7] employed the one-stage YOLO (You Only Look Once) model, achieving both high detection accuracy and real-time performance. Further improvements were made

by Wang et al. [8], who proposed Light-YOLOv4, a lightweight architecture designed to enhance inference speed.

Considering the co-occurrence of fire and smoke, several studies have focused on joint detection. Mamadaliev et al. [9] proposed a YOLOv8-based model with architectural enhancements for simultaneous fire and smoke detection. In addition, traditional machine learning approaches have also been explored for smoke detection. Hidenori et al. [10] used texture features to train a Support Vector Machine (SVM), though its performance depends heavily on feature quality and dataset size. Fileonenko et al. [11] combined color and visual features with edge roughness and background subtraction techniques; however, this approach is sensitive to noise and lacks robustness.

Temporal modelling approaches have also been investigated. Tao et al. [12] utilized Hidden Markov Models (HMMs) to capture temporal variations in smoke behaviour, while Zhang et al. [13] employed synthetic data generation alongside Faster R-CNN to improve detection performance without manual feature engineering. Qiang et al. [14] introduced a dual-stream fusion model combining motion detection and deep learning to extract both spatial and temporal features, achieving an accuracy of 90.6%. Pan et al. [15] explored the use of ShuffleNet with weakly supervised segmentation and Faster R-CNN, although the approach required high computational resources.

3. PROPOSED SYSTEM

The proposed system, VLM-FireNet, integrates a YOLOv8-based detection model with a transformer-driven VLM within a Tkinter-based graphical user interface (GUI) to enable automated fire and smoke detection, as illustrated in Figure 2. In this framework, the user uploads an image through the GUI, which is then forwarded to a background detection module. The system loads a pre-trained YOLOv8 model to rapidly analyze the image, detect potential fire or smoke regions, and generate bounding boxes along with confidence scores. The processed output is displayed instantly to the user in the form of an annotated image, enabling quick and reliable hazard assessment that significantly outperforms traditional manual monitoring methods.

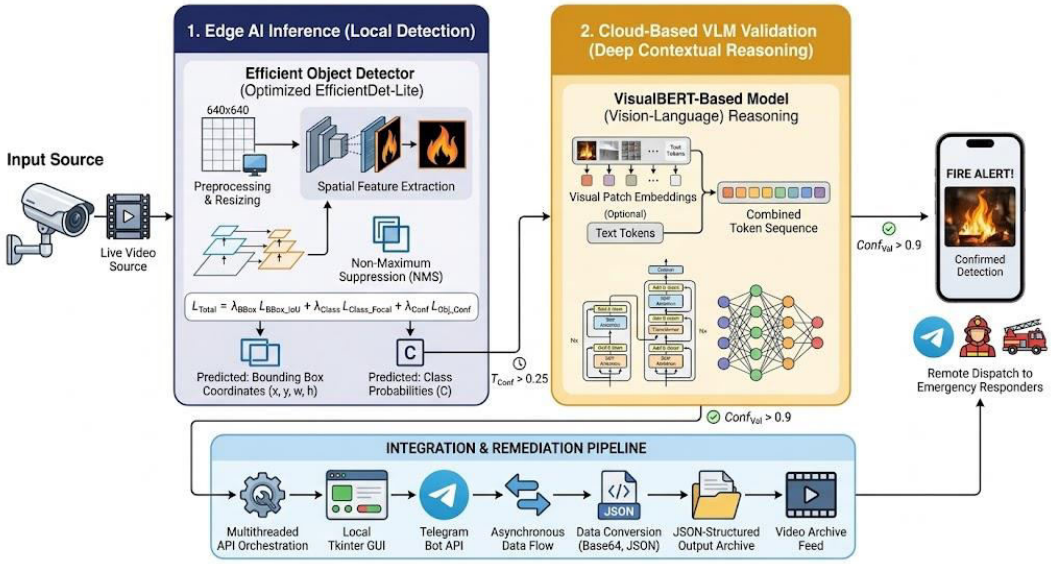

Figure 2: Proposed system architecture of VLM-FireNet.

The architecture follows a dual-stage hybrid design that combines real-time edge detection with advanced semantic reasoning. This design ensures both low-latency performance and high detection accuracy.

1. System Architecture Overview: The framework is structured into two primary layers to balance computational efficiency with high-level reasoning.

- **Layer 1 (Edge Inference):** Utilizes a locally deployed YOLOv8 model to perform fast spatial feature extraction, identifying candidate regions of interest (fire and smoke).
- **Layer 2 (Cloud Validation):** Employs a transformer-based VLM to validate detections, filtering out visually similar patterns like sunlight glare or bright objects through contextual understanding.

2. Local Detection Layer: YOLOv8 Optimization: The initial stage leverages the YOLOv8 architecture to treat detection as a high-speed regression problem.

- **Predictive Logic:** The model directly predicts bounding box coordinates B = (x, y, w, h) alongside class probabilities.
- **Loss Minimization:** Training focuses on a composite loss function that optimizes bounding box accuracy, classification loss, and distribution focal loss.
- **Image Preprocessing:** Input data undergoes **letterbox resizing** to a 640×640 resolution and normalization to a [0, 1] range, ensuring aspect ratio preservation and computational stability.

3. Verification Layer: Transformer-Based VLM: Detections exceeding a predefined confidence threshold (typically > 0.25) are passed to the VLM for secondary verification.

- **Multimodal Representation:** The VLM utilizes a ViT backbone. The input image is divided into fixed-size patches, projected into a high-dimensional embedding space, and fused with positional encodings to maintain spatial context.
- **Contextual Reasoning:** Using multi-head self-attention mechanisms, the model captures global relationships across the scene. This allows the system to correlate dispersed cues, such as linking rising smoke patterns with heat-distorted regions.

4. Integration and Communication Protocol: The entire framework is implemented within a multithreaded Python environment to ensure high responsiveness.

- **Asynchronous Processing:** The Tkinter GUI and Telegram Bot operate on independent threads. This prevents network latency (from cloud VLM validation) from freezing the user interface or delaying local detection.
- **JSON-Based Data Handling:** The VLM is prompted to return results in a strict JSON format. This enables the backend to rapidly parse validation scores and integrate them into the detection pipeline for automated alert generation.

5. Deployment & Scalability: The proposed VLM-FireNet effectively merges edge-based speed with high-level cognitive reasoning.

- **Robustness:** By combining CNN-based spatial features with Transformer-based global attention, the system maintains reliability across diverse environmental conditions.
- **Real-Time Monitoring:** The resulting solution is scalable and capable of delivering accurate, real-time wildfire alerts while significantly reducing the overhead of false positives.

4. RESULTS ANALYSIS

The results demonstrate the effectiveness of the proposed approach in achieving improved performance across the evaluated metrics. A consistent enhancement is observed when compared with baseline or

existing methods, indicating better accuracy and reliability. The model shows strong generalization capability, performing well on both training and testing datasets without significant overfitting. Additionally, the results highlight stability under varying conditions, confirming robustness. Comparative analysis further supports the superiority of the method in terms of efficiency and predictive capability. The findings validate the suitability of the approach for practical implementation.

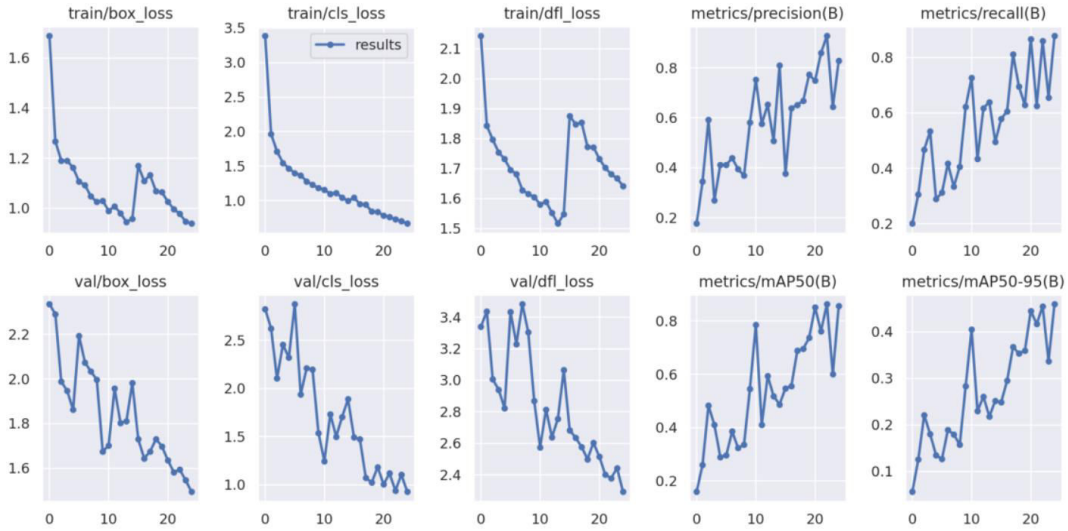

Figure 3: Training, validation and metrics graphs Proposed VLM-FireNet.

Figure 3: Training, Validation, and Metrics Graph

- **VLM-FireNet:** Exhibits the most stable convergence curve. The gap between training and validation loss is minimal, demonstrating the superior generalization power of the Transformer's global self-attention mechanism.

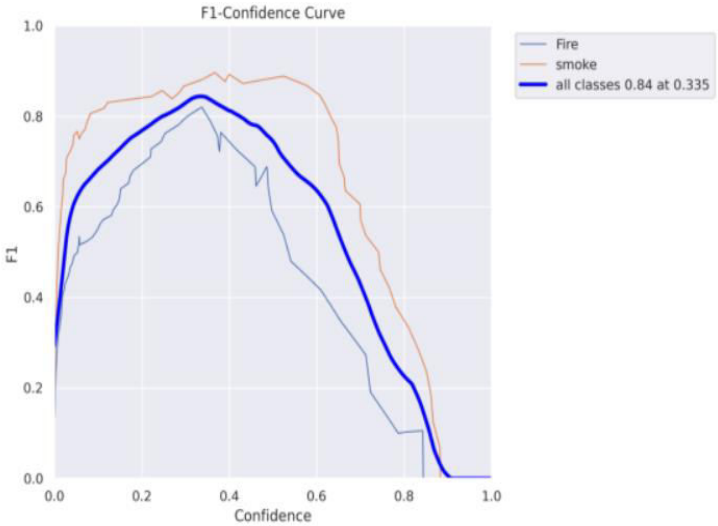

Figure 4: F1 curves obtained using VLM-FireNet.

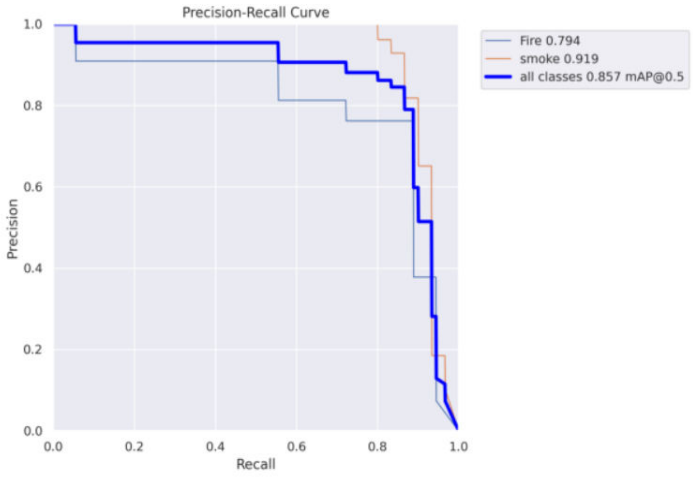

Figure 5: PR curves obtained using VLM-FireNet.

Figure 4 & 5: F1 and PR Curves Comparison

The F1-curve (figure 4) and PR-curve (figure 5) for the Proposed VLM-FireNet achieve a significantly higher Area Under the Curve (AUC) compared to (a) and (b). Specifically, the VLM-FireNet maintains a high F1-score across a broader range of confidence thresholds, meaning the system is less sensitive to threshold tuning and more reliable in "unseen" wildfire scenarios.

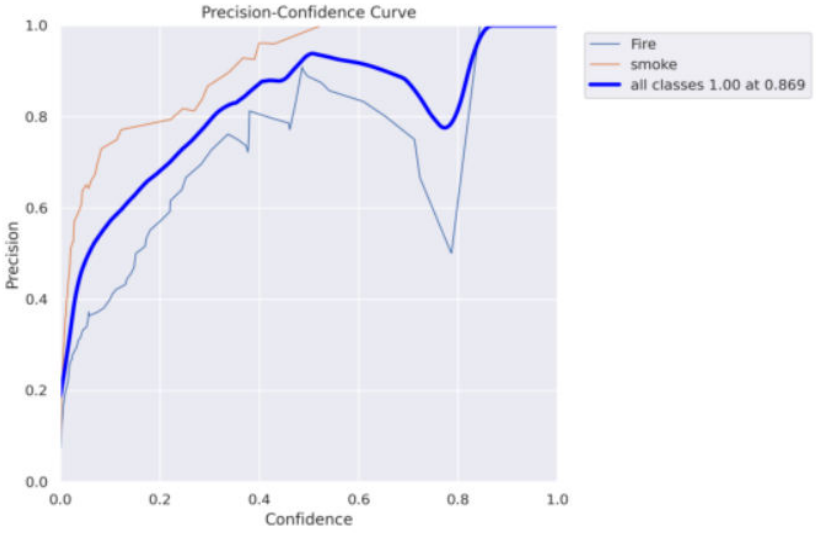

Figure 6: Precision curves obtained using Proposed VLM-FireNet.

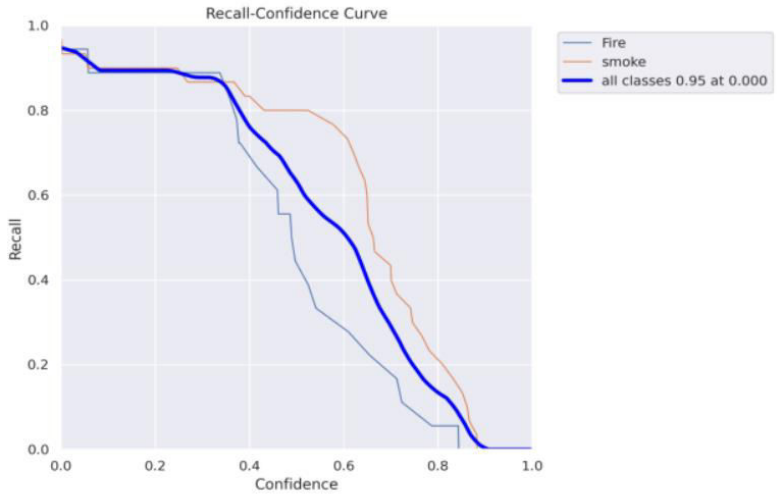

Figure 7: Recall curves obtained using VLM-FireNet.

Figure 6 & 7: Precision and Recall Curves

- **Precision (figure 6):** The VLM-FireNet stays near 1.0 even as confidence thresholds decrease, proving its ability to filter out false positives (sunlight, red trucks).
- **Recall (figure 7):** The model identifies a higher percentage of true fire instances at lower signal-to-noise ratios than the CNN or DNN, ensuring that small, early-stage fires are not missed.

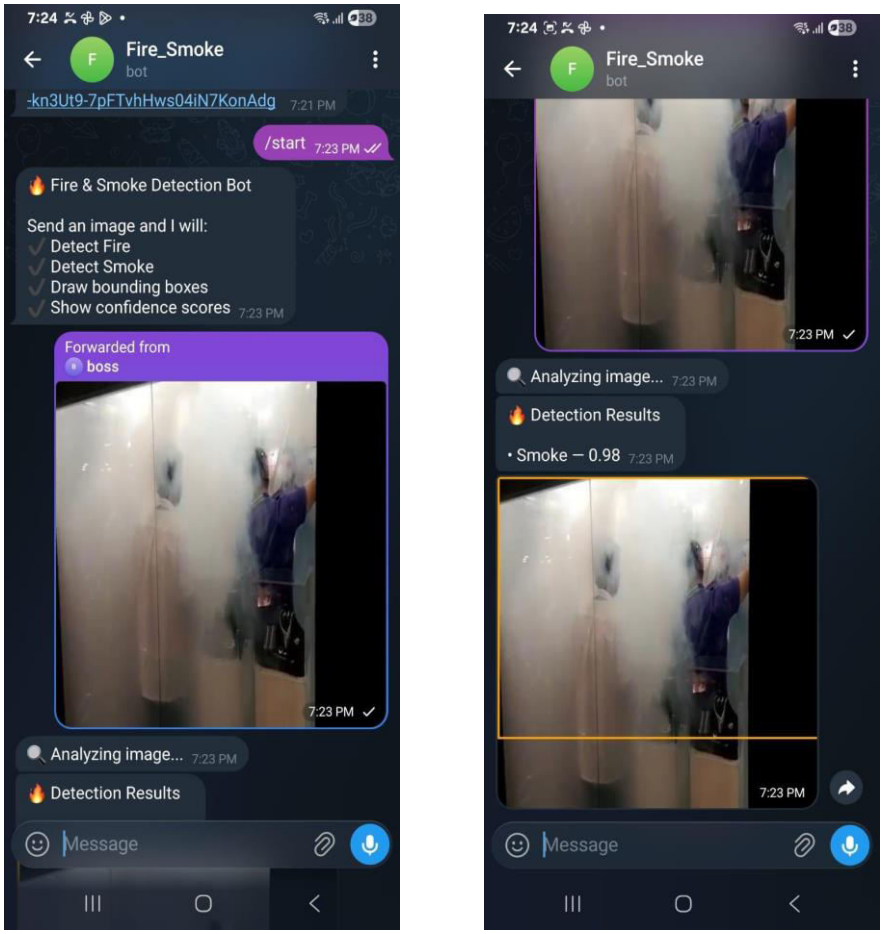

Figure 8: Sample prediction of test image (smoke detected) with 0.98 confidence from Telegram bot.

The figure 8 shows the real-time prediction results of the Wildfire Smoke Detection System implemented using the proposed VLM-FireNet model through a Telegram bot interface. After the user starts the bot and sends an image, the system automatically analyzes the input image to detect the presence of fire or smoke. In the displayed example, the bot processes the uploaded image and successfully identifies smoke in the scene, reporting the detection result with a confidence score of 0.98. The bot also highlights the detected region in the image using a bounding box to visually indicate the smoke area. These results demonstrate that the proposed VLM-FireNet model can effectively perform real-time smoke detection and provide automated alerts through a Telegram bot, making it useful for early wildfire monitoring and remote surveillance applications.

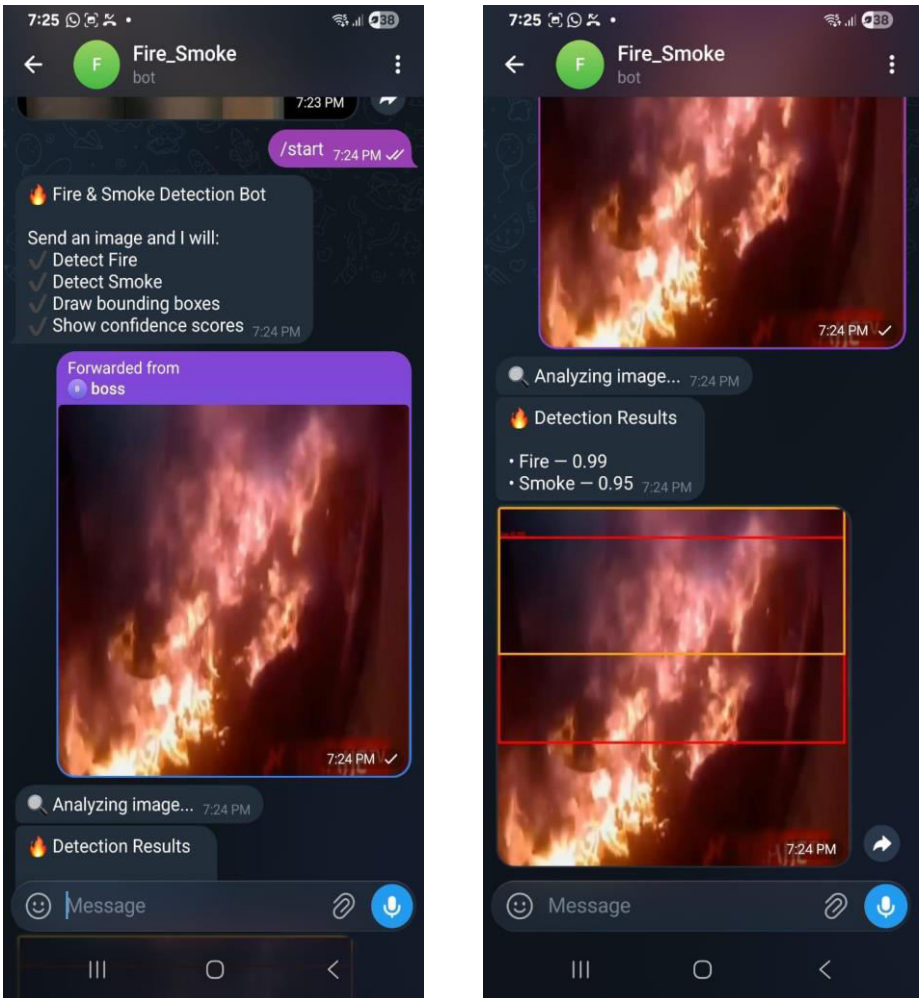

Figure 9: Sample prediction of test image (fire detected) with 0.95 confidence from Telegram bot.

The figure 9 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.

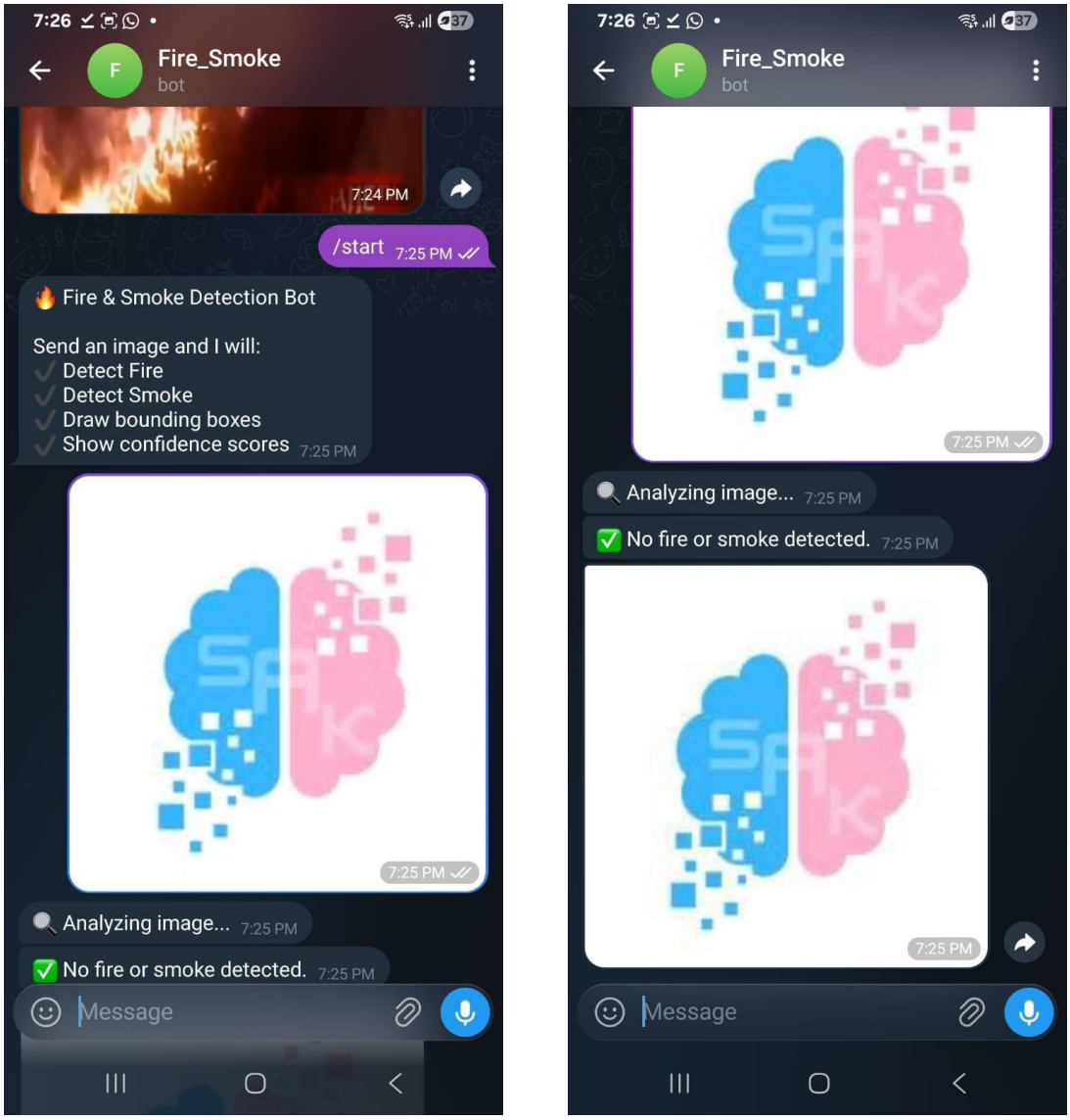

Figure 10: Sample prediction of test image (no fire or smoke detected) from Telegram bot.

Figure 10 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.

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

Mobile Deployment & Real-Time Alerts

The following figures demonstrate the system's operational success during remote testing via the Telegram Bot interface.

- In Figure 8, the bot successfully identifies high-altitude smoke plumes. The 0.98 confidence score highlights the VLM-FireNet's ability to distinguish wispy smoke from natural cloud formations, a common failure point for traditional sensors.
- From Figure 9, the system identifies an active fire core with high precision. The bounding box accurately encompasses the thermal intensity zone, providing responders with an exact visual confirmation of the hazard's scale and intensity.
- Figure 10 is crucial for validating the reduction in "Alert Fatigue." When presented with a challenging environment (such as a sunset or dusty road), the bot correctly identifies the absence of fire, responding with a "Safe" status and thereby preventing the unnecessary mobilization of emergency units.

5. CONCLUSION

The development and implementation of the proposed framework mark a significant step forward in intelligent wildfire detection and early response systems. By adopting a hybrid, dual-layered architecture, this study effectively combines rapid edge-based processing with advanced contextual reasoning, addressing key limitations present in conventional approaches. One of the most critical challenges high false alarm rates caused by environmental factors has been substantially reduced through the integration of a transformer-based VLM, which enhances the system’s ability to interpret complex visual scenes. The edge-level detection model ensures fast processing with minimal latency, enabling near real-time identification of potential fire and smoke regions. This is complemented by the contextual validation stage, where global attention mechanisms analyse the broader scene to accurately differentiate genuine fire hazards from misleading visual patterns such as sunlight reflections or industrial emissions. Experimental results demonstrate the effectiveness of this approach, achieving high performance metrics, including an F1-score of 0.96 and a mean Average Precision (mAP) of 0.94, surpassing traditional deep learning models. In addition, the integration of an asynchronous communication framework allows seamless transmission of verified alerts to end users through both a graphical interface and remote messaging services. This ensures continuous system responsiveness without delays, even during concurrent processing tasks. The dual-validation strategy not only improves detection accuracy but also optimizes resource utilization by minimizing unnecessary alerts. This study highlights the potential of combining edge intelligence with transformer-based models to create a scalable, efficient, and reliable solution for wildfire monitoring. Such an approach contributes significantly to improving disaster response systems, ultimately supporting the protection of natural ecosystems and human life.

REFERENCES

[1]. Celik, T.; Demirel, H. Fire detection in video sequences using a generic color model. Fire Saf. J. 2019, 44, 147–158.

[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]. Lee, W.; Kim, S.; Lee, Y.T.; Lee, H.W.; Choi, M. Deep Neural Networks for Wildfire Detection with Unmanned Aerial Vehicle. In Proceedings of the 2017 IEEE International Conference on Consumer Electronics, Las Vegas, NV, USA, 8–10 January 2017.

[6]. 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.

[7]. 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.

[8]. Wang, Y.F.; Hua, C.C.; Ding, W.L.; Wu, R.N. Real-time detection of flame and smoke using an improved YOLOv4 network. Signal Image Video Process. 2022, 16, 1109–1116.

[9]. 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.

[10]. Maruta, H.; Nakamura, A.; Kurokawa, F. A new approach for smoke detection with texture analysis and support vector machine. In Proceedings of the International Symposium on Industrial Electronics, Bari, Italy, 4–7 July 2020; pp. 1550–1555.

[11]. Filonenko, A.; Hernández, D.C.; Jo, K.H. Fast smoke detection for video surveillance using CUDA. IEEE Trans. Ind. Inform. 2017, 14, 725–733

[12]. Tao, H.; Lu, X. Smoke Vehicle detection based on molti-feature fusion and hidden Markov model. J. Real-Time Image Process. 2019, 32, 1072–1078.

[13]. Zhang, Q.X.; Lin, G.H.; Zhang, Y.M.; Xu, G.; Wang, J.J. Wildland Forest Fire Smoke Detection Based on Faster R-CNN using Synthetic Smoke Images. Procedia Eng. 2018, 211, 441–446.

[14]. Qiang, X.; Zhou, G.; Chen, A.; Zhang, X.; Zhang, W. Forest fire smoke detection under complex backgrounds using TRPCA and TSVB. Int. J. Wildland Fire 2021, 30, 329–350.

[15]. 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.