scientific reports

Check for updates

OPEN

Real time fire and smoke detection using vision transformers and spatiotemporal learning

Umesh Kumar Lilhore[1], Yogesh Kumar Sharma[2], Kavitha Venkatachari[3], Nikhil Kumar Jain[4], Sultan Aldossary[5], Shimaa A. Hussien[6], Ehab Seif Ghith[7], Lidia Gosy Tekeste[8]✉ & Sarita Simaiya[9]✉

The detection of fire and smoke in images and videos is essential for environmental monitoring and safety; however, the unpredictable nature of fires makes it a difficult task. Although traditional methods such as CNNs, LSTMs, and 3D-CNNs have made progress in fire detection, they frequently encounter difficulties in effectively integrating spatial and temporal information from both images and videos. In this study, we introduce a novel method that integrates Transformer attention mechanisms and Vision Transformers (ViTs) to enhance the precision of fire and smoke detection in both images and videos. ViTs are employed in our model to extract spatial features from images, leveraging their capacity to capture long-range dependencies, which are essential for the identification of fire and smoke. We utilise 3D-CNNs to extract spatiotemporal features from video sequences, while a Transformer encoder is used to track the evolution of fire and smoke over time. Furthermore, we execute various enhancements to optimise the model's performance. These encompass enhanced temporal modelling, advanced self-attention mechanisms, and a multi-task learning framework to improve the model's robustness by identifying potential hazards, such as smoke, fire, and other threats. In order to enhance the model's adaptability to dynamic environments, we incorporate sophisticated data augmentation techniques and optimize it for real-time deployment on edge devices. To address the inherent class imbalance between fire and non-fire samples in existing datasets, we implemented targeted data augmentation and class-weighted learning strategies, ensuring equal representation and balanced training for improved generalization. The model was tested against two well-known datasets: the NASA Space Apps Challenge Dataset and Kaggle's Fire Videos Dataset. Our method outperforms conventional methods, achieving 99.2% accuracy on the NASA dataset and 98.3% on the Fire Videos dataset. ResNet50, VGG16, LSTM, 3D-CNN, and hybrid ResNet50 + LSTM and VGG16 + 3D-CNN models, on the other hand, achieved accuracies ranging from 85% to 94%. This study's findings show that our hybrid model is a more effective solution for real-time fire and smoke detection in real-world settings because of its improved integration of spatial and temporal features.

Keywords Fire detection, Smoke detection, Vision transformers, 3D-CNNs, Multimodal fusion, Spatio-temporal features

Abbreviations

CNN	Convolutional neural network

[1]School of Computer Science and Engineering, Galgotias University, Gautam Buddh Nagar, Greater Noida 203201, Uttar Pradesh, India. [2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Field, Vaddeswaram, Guntur, AP, India. [3]School of AI & Future Technologies, Universal AI University, 410201 Karjat, Maharashtra, India. [4]Capgemini Technology Services India Limited, 139, 140 NSEZ, Phase-2, 201305 Noida, Uttar Pradesh, India. [5]Department of Computer Engineering and Information, College of Engineering in Wadi Alddawasir, Prince Sattam bin Abdulaziz University, Wadi Addawasir, Saudi Arabia. [6]Electrical Engineering Department, Faculty of Engineering, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia. [7]Department of Mechatronics, Faculty of Engineering, Ain Shams University, Cairo 11566, Egypt. [8]Eritrea Institute of Technology, Mai-Nefhi College, Mai Nefhi, Himbrti, Eritrea. [9]School of Computer Applications & Technology, Galgotias University, Gautam Buddh Nagar, Greater Noida 203201, Uttar Pradesh, India. ✉email: lidiagosytekeste@gmail.com; saritasimaiya@gmail.com

3D-CNN	Three-dimensional convolutional neural network
LSTM	Long short-term memory
SOTA	State of the art
YOLO	You only look once
RTX	Ray Tracing Texel eXtreme (NVIDIA GPU Series)
SDT	Smoke detection transformer
EfficientViT	Efficient vision transformer
IoT	Internet of things
FN	False negative
TN	True negative
BCE	Binary cross-entropy
GFLOPs	Giga floating-point operations
PR Curve	Precision–recall curve
NASA	National aeronautics and space administration
ViT	Vision transformer
FPS	Frames per second
FLOPs	Floating-point operations
AUC-ROC	Area under the receiver operating characteristic curve
GPU	Graphics processing unit
NX	NVIDIA Jetson Xavier NX
FireViTNet	Fire vision transformer network
MobileViT	Mobile vision transformer
FPS	Frames per second
FP	False positive
TP	True positive
AdamW	Adaptive moment estimation with weight decay optimizer
MB	Megabytes
ROC	Receiver operating characteristic curve
SNR	Signal-to-noise ratio

Fire and smoke detection is crucial for environmental protection and the safety of individuals in urban and rural settings. In urban environments, the swift identification of fire can markedly diminish economic losses linked to extensive fires, safeguard lives, and avert substantial infrastructural damage[1]. Public health and the environment are at risk when fires in buildings, industrial complexes, or forests escalate rapidly. Urban fire hazards, including building fires, have the potential to spread rapidly, affecting significant portions of a city's infrastructure and population[2,3]. In the same vein, wildfires are a perpetual threat in rural areas, threatening wildlife, consuming vast areas of forest, and contributing to environmental degradation through air pollution and the loss of biodiversity.

The necessity for sophisticated and dependable fire and smoke detection systems has become increasingly urgent as fire incidents continue to escalate on a global scale[4]. The capacity to identify fires in real time has the potential to reduce response times and improve public safety significantly. Similarly, the prompt identification of smoke is crucial for averting the escalation of fires, especially in expansive and complex settings like forests, airports, or industrial facilities. In these settings, the ecological consequences of the fire and the fatalities can be alleviated through prompt intervention and efficient detection[5].

The need for multimodal detection

Conventional fire detection systems typically depend on either image-based or video-based techniques to detect the presence of fire or smoke[6,7]. Nonetheless, each modality exhibits distinct limitations. Static images effectively capture clear visual indicators of fire; however, they do not offer insight into the temporal progression of fire behaviour. In contrast, video data provides a dynamic view of fire progression; however, it is frequently more challenging to analyse due to the large volume of data and the need to document temporal changes[8].

Multimodal fire and smoke detection, which combines image and video data, provides a more comprehensive solution[9]. Integrating spatial and temporal features from images and videos allows multimodal systems to improve detection accuracy and robustness. This method has the advantage of capturing both the visual characteristics and the progression of fire or smoke, allowing for more accurate and immediate detection[10]. For example, whereas an image may depict smoke ascending from a fire, video data can offer essential insights into the dispersion of that smoke, facilitating the forecasting of fire propagation and intensity. Multimodal systems represent a substantial improvement in fire detection, particularly in dynamic environments characterised by unpredictable fire behaviour[11].

Existing techniques for fire and smoke detection

Deep learning models have become increasingly important in the detection of fires and smoke. Convolutional Neural Networks are widely used for static image-based fire detection because of their ability to extract spatial features such as shape, texture, and colour that are unique to fire and smoke. Detecting fire in images is accomplished by these models, but they struggle to capture temporal information in video data[12,13]. Models such as 3D-CNNs and Long Short-Term Memory networks have been used for video-based fire detection to address this problem. LSTMs are ideal for capturing temporal dependencies across sequences of frames, whereas 3D-CNNs add a temporal dimension to traditional CNNs, allowing them to extract spatiotemporal features[14,15].

However, these models have their own set of limitations. While LSTMs are excellent at capturing long-term temporal dependencies, they frequently fail to capture fine-grained spatial features, which are required for accurate fire detection. 3D-CNNs, on the other hand, combine spatial and temporal features, but they are computationally expensive and necessitate a significant amount of memory and computing power[16,17]. This makes them less appropriate for real-time applications, particularly in environments with limited computational resources. Moreover, these models frequently encounter false positives and missed identifications in intricate, dynamic settings where fire or smoke may appear subtle or fluctuate rapidly between frames[18,19].

Alternative real-time detection models, including YOLO and Faster R-CNN, exhibit high efficiency in object detection tasks; however, they are primarily tailored for single-image analysis and neglect the temporal dimension of video data. These models may perform exceptionally well in static images; however, they encounter difficulties in monitoring the progression of fire or smoke over time, resulting in erroneous detection in video streams[20,21]. Additionally, U-Net models, which are frequently used for image segmentation, are highly effective in recognising fire in static images. However, they face difficulties when dealing with the dynamic characteristics of video data, which include the observation of fire and smoke as they develop in real-time[22,23].

Despite advances in deep learning models, real-world fire and smoke detection remains complex and dynamic. The challenge is to integrate spatial and temporal features efficiently while maintaining computational effectiveness. Furthermore, many models have poor adaptability to changing conditions, such as lighting, smoke density, or environmental factors, resulting in inaccurate detection. As a result, this highlights the necessity of developing more advanced hybrid models that are capable of overcoming these limitations and providing a method that is both more accurate and more efficient for detecting fire and smoke in images and videos[24,25].

Research motivations

Fire and smoke detection is crucial for protecting lives, securing property, and minimising environmental damage significantly, as the frequency and severity of fires increase due to climate change and urbanisation. Notwithstanding significant progress in fire detection via deep learning, existing methodologies still encounter difficulties in effectively tackling the complexities of real-world scenarios, particularly in the examination of images and videos[26,27]. Detecting fire and smoke in video streams necessitates the acquisition of both spatial and temporal information, a task that current models frequently execute ineffectively. The difficulty resides in merging spatial characteristics, such as flame configurations and smoke patterns, with temporal attributes, including the dynamics of fire propagation and smoke movement over time. Conventional models, such as CNNs, excel with static images but are inadequate for video analysis[28,29]. Conversely, models such as 3D-CNNs and LSTMs, which effectively process video, frequently incur substantial computational expenses, rendering them impractical for real-time applications. Furthermore, numerous models are optimised for particular environments, resulting in difficulties in generalising across varying lighting conditions, smoke densities, and fire behaviours[30,31].

This research is motivated by the necessity for a more resilient and efficient methodology for fire and smoke detection that functions effectively in both images and videos. The objective is to create a hybrid model that integrates ViTs for spatial feature extraction with Transformer attention mechanisms to monitor fire and smoke temporally. We aim to enhance accuracy, minimise false positives, and guarantee real-time performance, thereby addressing the principal limitations of current models. This work seeks to advance the reliability of fire detection systems, thereby improving public safety and facilitating informed decision-making in fire management, encompassing urban surveillance and forest fire monitoring[32,33].

Key contributions

The present research advances fire and smoke detection significantly by addressing critical challenges in spatial and temporal feature integration while also ensuring robustness and real-time functionality[34,35]. As a result, monitoring systems improve, resulting in more excellent safety and efficiency. The present investigation adds significant value to the field of smoke and fire detection by incorporating cutting-edge techniques to improve accuracy and real-time effectiveness. Following is a list of the different types of contributions.

- *Innovative Hybrid Model*: The present investigation proposes a novel hybrid model that combines 3D-CNNs for spatiotemporal feature extraction from video frames with ViTs for spatial feature extraction from images. By combining these models with transformer attention mechanisms, which efficiently capture both spatial and temporal dynamics, we can more accurately detect fire and smoke in images and videos.
- *Advanced Attention Mechanisms*: We improve the Transformer's self-attention mechanism to more effectively concentrate on critical fire and smoke attributes, thereby enhancing detection precision. This improved focus enables the model to differentiate between intricate situations involving fire, smoke, and various environmental factors, even in adverse conditions[4,36].
- *Multimodal Fusion for Robust Detection*: The model integrates various data types, such as enhancing detection abilities in low-visibility environments. The integration of multimodal data substantially improves the model's resilience and guarantees dependable fire detection across diverse environments.
- *Real-Time Optimisation for Edge Devices*: We enhance the model for instantaneous fire and smoke detection, rendering it appropriate for implementation on edge devices. Using sophisticated data augmentation techniques guarantees that the model generalises well in a variety of settings, allowing for quick and precise detection even on low-resource devices.
- *Balanced Learning for Imbalanced Data*: To mitigate the effects of class imbalance between fire and non-fire samples in the datasets, we employed targeted data augmentation, oversampling, and class-weighted loss functions. This ensures balanced learning, reduces false positives, and enhances the model's generalization across diverse real-world conditions[7,37].

- *Comprehensive Evaluation with Benchmarking*: We precisely evaluate the model using two prominent datasets: the NASA Space Apps Challenge Dataset and the Fire Videos Dataset from Kaggle. Our model outperforms established algorithms like ResNet50, VGG16, LSTM, 3D-CNN, and hybrid ResNet50 + LSTM, and VGG16 + 3D-CNN models.

Organisation of the article

The complete article is organised as follows. Section 2 begins with a Literature Review, which analyses the current methodologies for fire and smoke detection and identifies the deficiencies that our approach addresses. Section 3, Materials & Methods, details the model architecture, the datasets used, and the combination of Transformer attention mechanisms with Vision Transformers for multimodal detection. Section 4, Implementations & Results, outlines our model's experimental configuration, training specifics, and performance outcomes across diverse datasets and compares it to established methodologies. Section 5 concludes the article by summarising our findings and examining potential avenues for future research in fire detection.

Literature review

A wide range of approaches, from advanced deep learning models to conventional image processing techniques, are used in the extensive literature on fire and smoke detection. The selected works are divided into three primary categories for this review: Multimodal Fire Detection, Video-Based Fire and Smoke Detection, and Image-Based Fire Detection. Methods, datasets, significant contributions, limitations, and future research directions are highlighted in each category.

Image-based fire detection

Recent research has focused on fire detection in static images, using CNNs because of their ability to learn spatial patterns. Das et al. (2025) used the NV2CIR dataset to create an innovative attention network for identifying spliced video objects while also addressing privacy concerns about altered visual content. Although the proposed method successfully detects fire in altered images, it fails miserably in real-world, unaltered situations, primarily because of the complex environmental factors that influence fire detection. Zhang et al. (2024) presented a deep-learning framework for fire detection that uses CNNs and sensor data to predict backdrafts. Using the Fire Images Dataset, they concentrated on extracting essential features related to flame and smoke patterns; however, their model encountered difficulties in low-light conditions and with diminutive fire instances, leading to false negatives. Despite the efficacy of their framework in forecasting backdraught, further investigation is required to encompass a broader spectrum of environmental conditions.

Guo et al. (2024) proposed a method employing channel shuffling and adaptive spatial feature fusion to enhance the precision of fire and smoke detection. Their method showed notable enhancements in computational efficiency and detection accuracy in controlled settings, yet it is constrained in real-world scenarios with fluctuating fire intensities and smoke densities. Subsequent research should investigate additional optimisations for real-time, large-scale fire detection systems. Shahid et al. (2024) introduced a hybrid CNN-ViT architecture for fire recognition, integrating spatial and temporal features to enhance prediction accuracy. Their methodology, however, encountered difficulties with extensive data and exhibited insufficient generalisation when utilised for images depicting intricate fire scenarios. Their subsequent objectives concentrate on employing transfer learning to enhance model robustness and optimise training time efficiency. Li et al. (2022) devised a fire and smoke detection algorithm that uses a structured CNN network specifically designed for real-time applications. It was difficult for the algorithm to recognise smouldering fires or minor incidents with low-contrast features, despite the fact that it was successful in recognising major fire events. Their long-term research objectives centre on the incorporation of multi-scale feature extraction to enhance detection capabilities in complex environments.

Video-based fire and smoke detection

There have been encouraging results demonstrated by models that incorporate temporal information in video-based detection. In their presentation from 2024, Sun and Cheng introduced the Smoke Detection Transformer, which is a sophisticated real-time smoke detection model that is designed to provide timely fire notifications. Using Transformers, they were able to successfully capture temporal dependencies in surveillance videos, which resulted in an improvement in detection capacity in comparison to conventional CNNs. Nonetheless, their methodology encountered issues with slower processing speeds when dealing with high-resolution video streams, necessitating additional optimisation for real-time applications.

Yang et al. (2024) presented an improved iteration of the RT-DETR framework intended for real-time smoke detection in surveillance footage. To improve the detection of smoke pattern evolution, the model combined triplet attention with a hierarchical feature pyramid network (HS-FPN). The model performed well in dynamic environments but had difficulty identifying objects in intricate indoor settings with fluctuating lighting conditions, highlighting the need for improved illumination management in future advancements. A deep learning model with 3D convolutional layers and attention mechanisms was employed by Guo et al. (2024) to segment fire and smoke in videos. Their channel-shuffling technique was capable of detecting fire and smoke in challenging environments; however, it encountered difficulty in distinguishing smoke from fog or other environmental elements, which could result in misclassifications. This could be enhanced in the future by incorporating supplementary contextual attributes to facilitate more accurate ecological differentiation.

The FireViTNet model, which employs CNNs and ViTs to segment forest fires, was introduced by Wang et al. (2024). Their model demonstrated exceptional precision in the detection of fire in video feeds obtained from remote sensors, particularly in wooded areas. Nonetheless, the model's efficacy was found to be reduced

in urban settings, where fire patterns and backgrounds are more variable. The authors propose future enhancements to urban fire detection by increasing the model's robustness across diverse environments. Shahid et al. (2023) used LSTMs and Transformer networks to track fire progression and detect smoke in video sequences. They demonstrated that the combination of these models could better capture the temporal dynamics of fire propagation. Nonetheless, their model encountered limitations when addressing occlusions or partial perspectives of fire, a common challenge in surveillance footage, implying that future research should focus on resolving occlusion-related difficulties in video-based fire detection.

Multimodal fire and smoke detection

The integration of image and video data into a multimodal framework is a developing field in the field of fire and smoke detection research. Mowla et al. (2024) developed a Hierarchical Multi-Headed CNN for the detection of aerial forest fires. This CNN incorporates data from both visual and environmental sensors. However, their hybrid model encountered difficulties with ecological noise, particularly in regions with dense fog or smoke, where the sensor data was unable to complement the visual data effectively. In order to enhance the accuracy of models under challenging environments, future research should investigate more robust sensor fusion techniques.

Yar et al. (2024) proposed a multimodal approach for fire detection in intelligent surveillance systems that combines Transformer networks and CNNs. Their method effectively integrated both spatial and temporal features, resulting in satisfactory performance on the Fire Videos Dataset. Nevertheless, their model was less effective in identifying small-scale fire incidents or those that took place in highly complex environments. Additional research is required to enhance the sensitivity of detection in densely populated or highly cluttered environments. RFWNet, a multi-scale remote sensing network that employs both video and image data for wildfire detection, was introduced by Wang et al. (2024). Their model was especially effective at detecting large wildfires; however, it struggled to detect smaller fires or those in urban areas, where fire patterns may deviate from typical large-scale characteristics. The authors suggest that future research should use adaptive learning techniques and higher-resolution sensors to manage more complex and smaller fires.

Chen et al. (2024) presented a lightweight fire hazard recognition model tailored to urban subterranean buildings. This model detects fire hazards by utilising sensor and image data. The model's capacity to generalise to real-world scenarios was impaired by differences in smoke properties and sensor noise despite its successful performance in controlled environments. In order to improve the robustness of models in unpredictable environments, future research would benefit from more diverse training data and improved sensor calibration. Gragnaniello et al. (2024) conducted a thorough examination of fire and smoke detection methods, classifying a variety of techniques under a novel taxonomy. Shan et al. (2021) introduced DRRNets (Dynamic Recurrent Routing via Low-Rank Regularisation in Recurrent Neural Networks), which combine dynamic routing and low-rank regularisation to improve the efficiency of RNNs for sequential tasks. This approach optimises information flow and reduces model complexity, enhancing the ability to capture complex temporal dependencies while preventing overfitting. DRRNets offers a promising method for improving performance in tasks involving sequential data, such as video-based fire and smoke detection.

The advantages of employing deep learning models such as Transformers and CNNs for both image and video-based detection were emphasised in their research. Nevertheless, they acknowledged the difficulty of effectively integrating multimodal data in real-time systems. Their review indicates that future advancements should concentrate on fusion techniques that can seamlessly integrate various types of data to facilitate more precise and timely fire detection. Table 1 presents a comparative analysis of various existing research in fire detection.

Materials & methods

This section outlines the datasets, model architecture, and various methodologies utilised to develop and evaluate the proposed hybrid fire and smoke detection model. The proposed method incorporates ViTs, 3D-CNNs, Transformer attention mechanisms, and a multi-scale fusion strategy to address the challenges of detecting fire and smoke in images and videos.

Dataset description

In the present investigation, we assessed the proposed fire and smoke detection model's effectiveness using two significant datasets: the NASA Space Apps Challenge Dataset and the Kaggle Fire Videos Dataset. These datasets are widely used in fire detection research because they provide a variety of scenarios for evaluating models in both static image and dynamic video formats, which are necessary for practical applications.

NASA space apps challenge dataset

There are 999 images in the processing, 755 of which are of fire and 244 of which are not. These photos provide a thorough dataset for training and evaluating models for identifying fire-related situations since they show smoke and fire in a variety of outdoor settings[33].

- *Fire Images*: This dataset includes images of a variety of outdoor fire scenarios, such as forest fires, wildfires, and fires with varying levels of smoke. Specific images depict large, vibrant flames, while others depict smaller fires. Some images emphasise dense smoke that is devoid of visible flames. This variety is indispensable for the development of a model that is capable of identifying fire at different stages of a fire event.
- *Non-Fire Images*: The non-fire images depict natural landscapes, including forests, rivers, lakes, and foggy regions. These images must be used to instruct the model on how to differentiate between fire-related images

Author(s)	Method(s)	Dataset(s)	Contribution	Outcome	Multimodal	Limitation(s)	Future work
Elhanashi et al. (2025)	Deep learning review	Various datasets	Comprehensive review of models, datasets, and challenges	High-level insights into fire and smoke detection	Yes	No unified model or framework	Propose unified models for future research
Khan et al. (2025)	Deep learning surveillance feeds	Indoor & outdoor feeds	Advancements in fire and smoke detection for surveillance	Significant improvement in real-time surveillance	Yes	Limited to surveillance feeds	Extend to dynamic real-time environments
Chen et al. (2025)	Multi-task learning, CNN	Forest fire images	Efficient multi-task forest fire and smoke detection model	High accuracy in forest fire detection	No	Struggles with noise in complex environments	Improve noise resilience
Panindre et al. (2025)	AI-integrated IoT, autonomous systems	Live video streams	Real-time fire and smoke detection via IoT	Real-time remote monitoring for fire detection	Yes	Real-time video processing limitations	Improve edge device efficiency
Das et al. (2025)	Attention networks	NV2CIR dataset	Attention network for detection of spliced video objects	Improved object detection and privacy issues	Yes	Limited dataset for generalisation	Expand the dataset for better generalisation
Zhang et al. (2024)	CNN, sensor fusion	Fire images	Fire detection with backdraft forecasting	High accuracy, weak in low light & small fires	No	Low-light, minor fire issues	Improve detection
Guo et al. (2024)	Channel shuffling, spatial fusion	Fire & smoke	Hybrid fire detection model	Accurate, efficient	No	Struggled with varied intensities	Enhance real-world use
Sun & Cheng (2024)	Transformer networks	Fire video	Real-time smoke detection	High accuracy, slow on high-res video	No	Processing speed	Optimize speed
Yang et al. (2024)	RT-DETR, triplet attention	Surveillance video	Real-time smoke detection	Improved detection, weak indoors	No	Indoor settings	Enhance indoor accuracy
Guo et al. (2024)	3D-CNN, attention mechanism	Fire video	Spatio-temporal fire detection	High accuracy, misclassified fog	No	Struggles with fog	Improve differentiation
Proposed hybrid model (2025)	ViT + 3D-CNNs + Transformer attention	NASA space apps challenge, fire videos	Hybrid approach integrating ViTs, 3D-CNNs, and Transformers	High accuracy (99.2% NASA, 98.3% fire videos), real-time detection	Yes	Computational complexity (large model size, inference time)	Enhance scalability for real-time applications

Table 1. Comparative analysis of various existing research in fire detection.

and typical outdoor scenes. The dataset encompasses a diverse array of environmental settings, guaranteeing that the model is exposed to a variety of backgrounds and contexts.

Fire videos dataset
Kaggle's Fire Videos Dataset contains video clips of fire incidents in various environments. This dataset is critical for testing fire detection models on temporal data because it enables the detection of fire and smoke at various stages of a fire event and over time[35].

- *Diverse Environments*: This dataset contains videos of fire incidents that occurred in various environments, including forests, industrial buildings, and open fields. These settings present unique challenges to fire detection models. For example, smoke in forest fires can obscure the flames, whereas fires in buildings can move much faster, forcing the model to adapt to rapid changes in the video.
- *Temporal Dynamics*: Unlike static images, video data consists of temporal sequences of frames. The Fire Videos Dataset allows you to see how well the model tracks fire and smoke over time. This is critical for understanding the dynamics of fire spread, smoke dispersion, and intensity changes. The temporal aspect of video data complicates the model because it must account for not only the appearance of the fire at any given time but also how it evolves.

Data pre-processing

Both the NASA Space Apps Challenge Dataset and the Fire Videos Dataset were prepared for training and evaluation by utilising a series of crucial data pre-processing steps. These steps were used for the fire and smoke detection model. When these pre-processing steps were carried out, the primary objective was to guarantee that the data were consistent, to strengthen the model, and to improve its capacity to generalise to a variety of environmental conditions. In the course of the data pre-processing procedure, the following are the specific steps that are carried out:

Dataset handling
The following key steps were used for dataset handling.

- NASA Space Apps Challenge Dataset: This dataset consists of 999 images, including 755 images containing fire and smoke and 244 non-fire images. For each image, we executed the subsequent pre-processing operations: resizing, normalisation, and augmentation (Fig. 1 presents the Sample Fire and Smoke Images in the dataset).
- Fire Videos Dataset: The Fire Videos Dataset comprises video clips depicting scenes related to fires captured in various settings, including open fields, buildings, and forests. We extracted frames from each video at predetermined intervals to ensure that the model received consistently accurate input data.

Fig. 1. Sample fire and smoke images in the dataset[33,35].

Image preprocessing for static image
The following key operations were performed.

- *Resizing*: All images were changed to be the same size, 224 by 224 pixels, which is a standard size used in deep learning models like CNNs and Vision Transformers. Consistency in input sizes is essential for the effective training of the model, and this standardisation ensures this[36,38].
- *Normalisation*: The pixel intensity values of the images were normalised to a range of [0, 1]. In order to accomplish this, the value of each pixel was divided by 255. This was done because the intensity values of pixels in an 8-bit image range from 0 to 255[31]. From a mathematical standpoint, this can be expressed as presented by Eq. 1. Here $\breve{X}_{i,j}$: Normalised pixel value, $X_{i,j}$: Original value.

$$\breve{X}_{i,j} = \frac{X_{i,j}}{255} \tag{1}$$

Video frame preprocessing (for the fire videos dataset)
The following key operations were performed.

- Frame Extraction: For each video, frames were extracted at regular intervals. We sampled one frame per second to generate a series of frames depicting the fire event chronologically. This method guarantees that the model accurately reflects dynamic alterations in the fire situation[32].
- Resizing and Normalisation: Each frame was resized to a uniform size of (224×224) pixels and normalised by dividing the pixel values by 255 to scale them between 0 and 1, as described in the image preprocessing step.
- Augmentation: We applied the same image enhancement techniques that we use for images to every video frame. These techniques included random rotation, flipping, cropping, and lighting modifications. This improvement ensures that the model is able to accommodate variations in the camera angle, lighting, and other environmental conditions that are present in the videos[33].
- Temporal Sequence Construction: We grouped consecutive frames to form temporal sequences, as the Fire VideosDataset is composed of sequences of frames. By utilising a temporal window size of 10 frames, the model was ableto capture the dynamics of fire and smoke progression over time.

Data augmentation, dataset balancing, and weighted loss function
In order to increase the robustness of the model and counter overfitting, an extensive data augmentation and balancing pipeline was employed. Various augmentation techniques, summarized in Table 2, were utilized to model different environmental scenarios and camera angles, which are generally faced by fire and smoke detection problems. Such operations help the model to generalize well for various real-world conditions (i.e., changing illumination, haze, and reflection as well as the pattern of dense smoke)[37,39]. Thus, data augmentation was strategically used not only for expanding dataset diversity but also to overcome class imbalance between fire and non-fire samples in both the NASA Space Apps Challenge and Kaggle Fire Videos datasets. The original mixes of examples were 3:1 in favour of fire images and frames. Such an imbalance may potentially lead to the

Augmentation technique	Application purpose	Description
Random rotation	Balancing & generalization	Rotates images randomly by ± 15°–30° to simulate varying orientations of fire and non-fire scenes.
Random horizontal and vertical flip	Balancing & generalization	Flips images along both axes to replicate different camera perspectives and viewpoints.
Random cropping	Balancing	Crops specific regions of images to simulate zoom effects and focus variability in fire and non-fire contexts.
Brightness, contrast, and saturation adjustments	Balancing & generalization	Modifies lighting and color intensity to mimic variations such as fog, smoke haze, sunlight, or artificial illumination.

Table 2. Data augmentation techniques used for balancing and generalization.

Dataset	Stage	Fire count	Non-fire count	Total count
NASA space apps challenge	Before preprocessing	755	244	999
	After preprocessing (Resizing + Normalization)	755	244	999
	After balancing (Targeted augmentation)	755	755	1,510
Kaggle fire videos	Before preprocessing	12,000	4,800	16,800
	After preprocessing (Frame extraction + Normalization)	12,000	4,800	16,800
	After balancing (Targeted augmentation)	12,000	12,000	24,000
Combined dataset	–	12,755	12,755	25,510

Table 3. Dataset details before and after preprocessing, augmentation, and balancing.

classifier being biased towards classifying samples as the majority class, resulting in a high rate of false positives in real-world use.

To address this, focused oversampling and augmentation were performed only on the minority (non-fire) class. Transformations, including rotation, flipping, random cropping, and controlled brightness-contrast-saturation variations, were applied to create realistic non-fire cases that simulate difficult natural situations (e.g., foggy, luminous dawn, low light, or overcast conditions). The minority class was stretched by these transformations without breaking the semantics enabling us to achieve a 1:1 class distribution in all datasets. In addition, class-weighted binary cross-entropy (CW-BCE) loss was employed for algorithmic-level fairness under training. This function emphasizes the minority class by assigning higher weights, and maintains balanced gradient updates for false fire detection and missed fire detection. The weighted loss is as follows (Eq. 2).

$$L = -[w_1\, y \log(p) + w_0\, (1-y) \log(1-p)] \tag{2}$$

Where: $y \in \{0,1\}$ denotes the ground truth label (1 = fire, 0 = non-fire), p is the predicted probability of fire, w_1 and w_0 represent the weights for fire and non-fire classes, respectively, calculated as (Eq. 3).

$$w_1 = \frac{N}{2N_1}, w_0 = \frac{N}{2N_0} \tag{3}$$

Here, N_1 = number of fire samples, N_0 = number of non-fire samples, $N = N_1 + N_0$ = total number of samples.

This weighting ensures that the model penalizes false fire and missed non-fire cases equally, thereby maintaining a balanced trade-off between precision and recall.

After augmentation and balancing, the final dataset comprised 25,510 total samples, evenly distributed between 12,755 fire and 12,755 non-fire instances. This comprehensive preprocessing pipeline combining targeted augmentation, class balancing, and weighted loss ensured that the proposed hybrid model learned from a fair, diverse, and representative dataset, reducing bias toward the fire-dominant class and significantly enhancing its ability to perform reliable real-time detection under diverse environmental and visual conditions (Table 3).

Impact of augmentation on video temporal coherence

While data augmentation enhances model robustness and reduces overfitting, it is essential to preserve temporal coherence when applied to video data. Since videos inherently contain sequential dependencies between frames, improper augmentation can disrupt motion continuity and mislead temporal feature extraction. To address this, all augmentation operations were carefully designed to maintain temporal consistency across video sequences:

- Uniform Frame-Level Transformation: Each augmentation (rotation, flip, crop, brightness, and contrast adjustments) was applied identically to all frames within a video sequence, ensuring consistent spatial alignment and preventing temporal distortion.
- Consistent Flipping and Rotation: Random flips and rotations were executed with identical parameters across frames, preserving spatial relationships of fire and smoke movement over time.
- Controlled Cropping: The same cropping window was maintained across all frames, ensuring that spatial focus and frame alignment remained stable throughout the sequence.

- Frame Extraction Strategy: Frames were sampled at a fixed rate of one frame per second to capture the dynamic progression of fire and smoke without temporal gaps or redundancy.
- Uniform Lighting Adjustments: Brightness, contrast, and saturation modifications were consistently applied to entire sequences, ensuring uniform illumination across frames and preserving the perceived motion of fire and smoke.

By maintaining these constraints, the temporal coherence of the video data was fully preserved. This allowed the model to effectively learn spatiotemporal dynamics capturing both the evolution and motion patterns of fire and smoke without compromising sequential integrity or introducing artificial inconsistencies.

Proposed model architecture

The proposed hybrid model integrates Vision Transformers and 3D-CNNs augmented by a Transformer encoder to proficiently capture spatial and temporal dependencies for fire and smoke detection in images and videos. The architecture is crafted to utilise the advantages of each component to tackle the intricacies of fire and smoke detection in various conditions. Figure 2 presents the Architecture for the Proposed Hybrid Model. The complete working of the proposed hybrid model is as follows.

Input dataset

This research's input dataset comprises two distinct but complementary sources: the NASA Space Apps Challenge Dataset and the Fire Videos Dataset from Kaggle. These datasets were meticulously selected to encompass a broad spectrum of fire and smoke scenarios in both static images and video sequences.

Vision transformer

The ViT is fundamental to the model's capacity to extract spatial features from static images in the proposed fire and smoke detection system. In contrast to CNNs, which employ convolutional layers to extract spatial features hierarchically, ViTs regard images as sequences of patches and utilise self-attention mechanisms to discern relationships among these patches[36,15].

The primary benefit of employing ViTs in this model is their capacity to capture long-range dependencies within images, rendering them exceptionally proficient in detecting fire and smoke over extensive regions. Fire and smoke, particularly in intricate outdoor settings, frequently occupy substantial areas of an image. By segmenting the image into patches and concurrently processing these patches via self-attention, ViTs can more efficiently acquire contextual relationships, even when these features are dispersed throughout the image[17]. Figure 3 presents the Architecture of the Vision Transformer, and Algorithm 1 presents the working steps.

3D-CNN for video frames

The proposed model employs 3D Convolutional Neural Networks (3D-CNNs) to extract spatiotemporal features from video data, which is crucial for capturing the dynamic attributes of fire and smoke propagation across consecutive video frames. CNNs are designed to analyse spatial features from an individual image[16]. However, 3D-CNNs improve this capability by incorporating a temporal dimension. This enables the model to recognise spatial patterns within individual frames and temporal dependencies that demonstrate the evolution of fire and smoke over time.

The potential of 3D-CNNs to depict the progression of fire and smoke over a series of frames is the primary advantage of using these networks in the detection of fire and smoke. It is insufficient to identify fire or smoke in a single frame in video-based detection; the model must monitor the progression of fire and smoke over time, taking into account its evolution across successive frames[10,26]. 3D-CNNs achieve this by employing convolutional filters in both spatial and temporal dimensions, which enables the model to understand the interaction of fire and smoke over a variety of moments in the video sequence and to discern motion patterns. Figure 4 presents the Architecture of a 3D-CNN for Video Frames, and Algorithm 2 presents the working steps.

Fusion layer

The Fusion Layer is critical in the proposed hybrid model because it combines spatial features extracted from static images using the ViT and spatiotemporal features extracted from video frames using the 3D-CNN. This layer is critical because it combines insights from both modalities (image and video) to improve the overall efficiency of fire and smoke detection. The proposed methodology employs a multi-scale feature fusion strategy that amalgamates features from two distinct models (ViT for images and 3D-CNN for video) into a cohesive representation. This integration enhances the model's comprehension of fire and smoke detection dynamics across various data types[2,11].

The Fusion Layer is in charge of aligning the spatial features from ViT with the temporal features from the 3D-CNN. This integration enables the model to effectively use both types of information, static features like color, texture, and shape from images, and dynamic temporal patterns like fire propagation and smoke movement from video[14]. Figure 5 presents the architecture, and Algorithm 3 presents the working steps.

Transformer encoder with self-attention method

The Transformer Encoder is a crucial component of our proposed hybrid model, designed to capture temporal dependencies in video sequences. Processing the spatiotemporal features extracted from video frames by the 3D-CNN is essential. The Transformer Encoder's ability to capture long-term dependencies and temporal variations makes it especially effective for fire and smoke detection in videos, where the progression of fire and the dispersion of smoke are essential for accurate identification[23,27]. Figure 6 presents the architecture, and Algorithm 4 presents the work of the Transformer Encoder with the Self Attention Method.

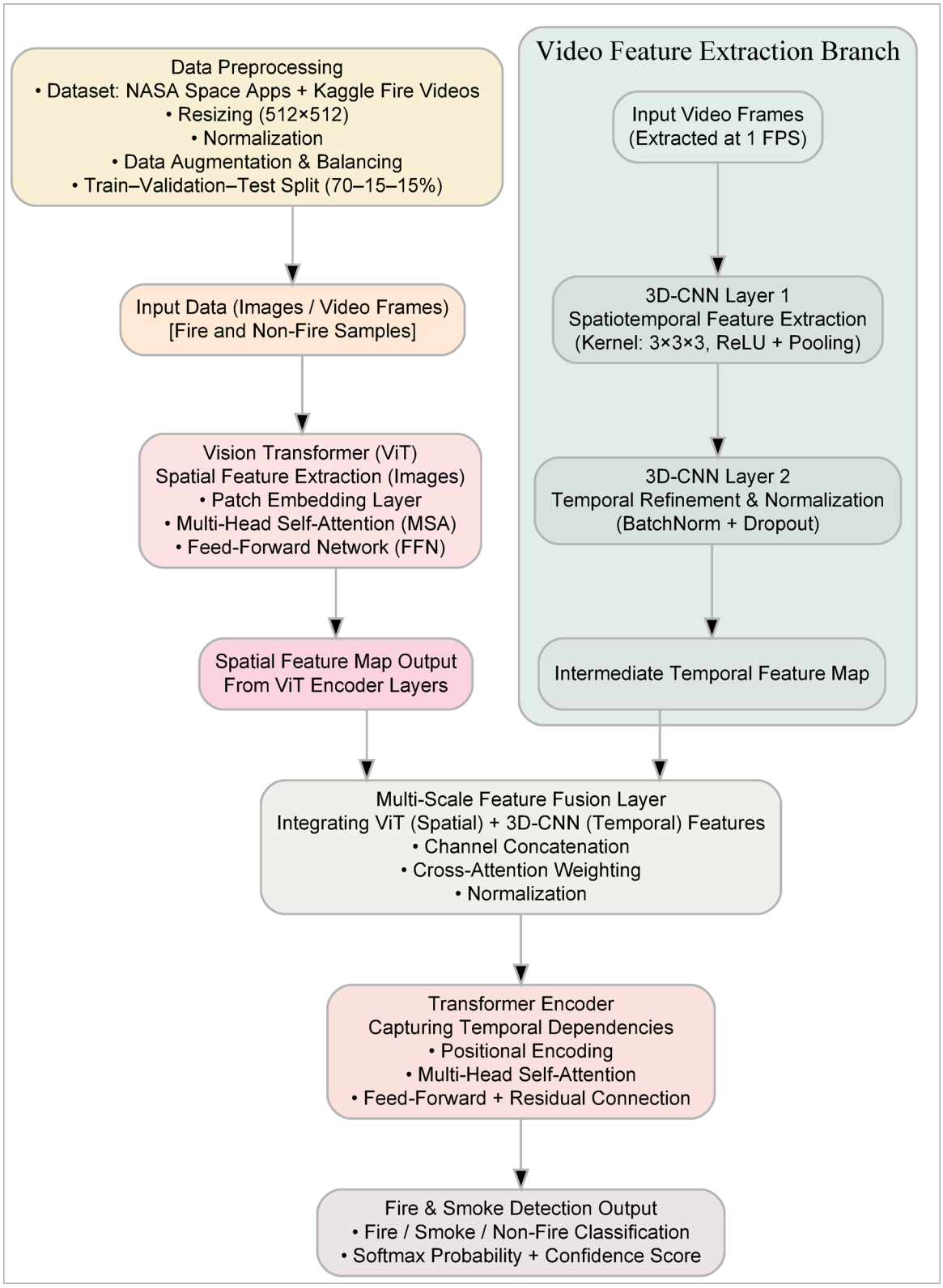

Fig. 2. Architecture of proposed hybrid transformer–3D-CNN model for fire and smoke detection.

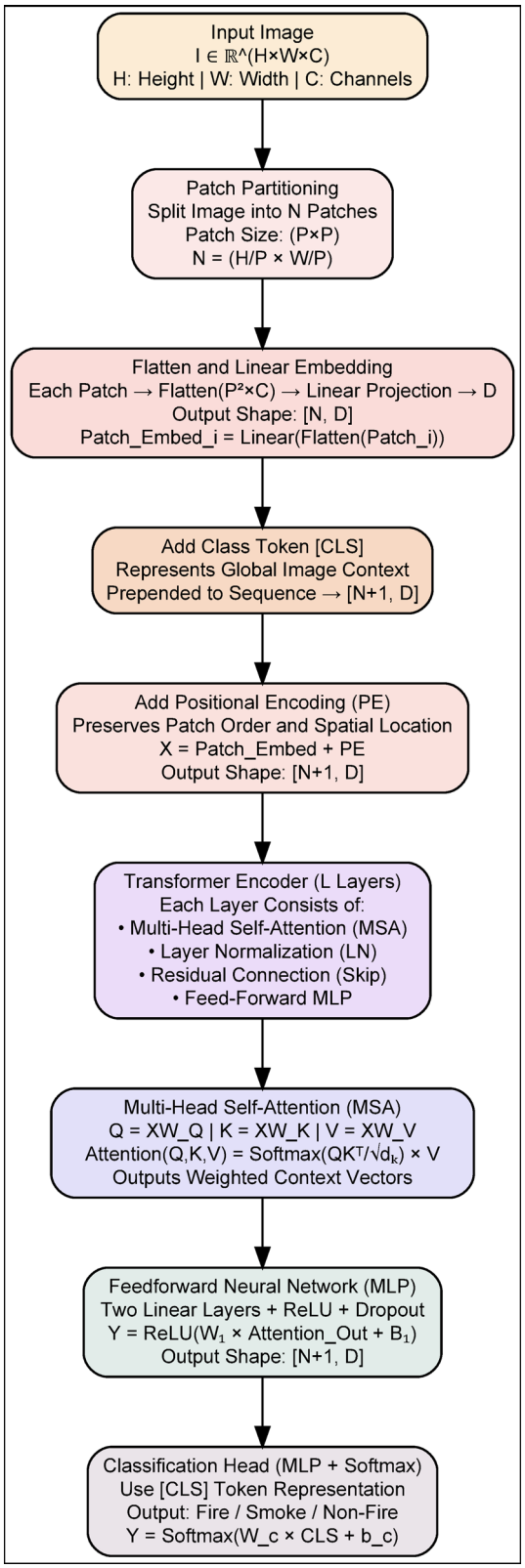

Fig. 3. Architecture of vision transformer (ViT) in the proposed hybrid model.

Input: Image $I \epsilon R^{(H*W*C)}$

Output: The ViT generates a classification output (fire or no fire) for the image derived from the acquired features.

Step 1: Input image dataset

1.1. Take an input image $I \epsilon R^{(H*W*C)}$*, Where H: Height, W: Weight, and C: Channels*

Step 2: Patch Partitioning

*2.1 The image has been divided up into various patches of size (P*P) with differing densities, Equations 4 and 5.*

$$N = \left(\frac{H}{P} \times \frac{W}{P}\right) \quad (4)$$

$$Patch_i = I[i: i + P, j: j + P] \quad (5)$$

Here $(i, j) \epsilon \{0, P, 2P, \ldots \ldots nP\}$*, P: Patch size, N: Number of patches,*

Step 3: Flattening

3.1 Convert each patch into a vector by flattening it by Equation 6.

$$Flattendedpatch_i = Flatten(patch_i) \epsilon R^{P^2 * C} \quad (6)$$

Here C: Chanel, R: Real Numbers,

Step 4: Linear Projection

4.1 For the purpose of projecting the patches into a space with a higher dimension, each flattened patch is passed through a linear layer, Equation 7.

$$Patch_{Embeddin\ g_i} = Linear(Flattendedpatch_i) \epsilon R^D \quad (7)$$

Here D: embedding dimension, D: Embedding dimension (output dimension of the linear projection).

Step 5: Positional Encoding.

5.1 Patch embeddings that contain positional encodings are subjected to multiple layers of self-attention, including the following components presented by Equation 8, calculated for each patch embedding.

$$Position_{Encodin\ g_i} = PE_i \quad (8)$$

Here PE_i*: Positional encoding for the* i^{th} *patch.*

Step 6: Self-Attention

6.1 It is necessary to pass the patch embeddings that contain positional encodings through multiple layers of self-attention, Equation 9.

$$Attention\ (Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right) * V \quad (9)$$

Here Q: Query, K: Key, V: Value Matrices.

Step 7: Feedforward Layer

The output of the self-attention is then transported through a feedforward neural network, which is represented by Equation 10.

$$MLP_{Output} = ReLU(W_1 * Attention_{Output} + B_1 \quad (10)$$

Here B: Bias value, W: Weight, $ReLU$ *: Rectified Linear Unit activation function,* MLP_{Output} *: Output of the feedforward neural network (MLP)*

Step 8: Final Output

8.1 A class prediction (fire or no fire) for the image is what the ViT produces as its output. This prediction is based on the features that were learned.

Algorithm 1. Algorithm for working of vision transformer (ViT) in a proposed hybrid model.

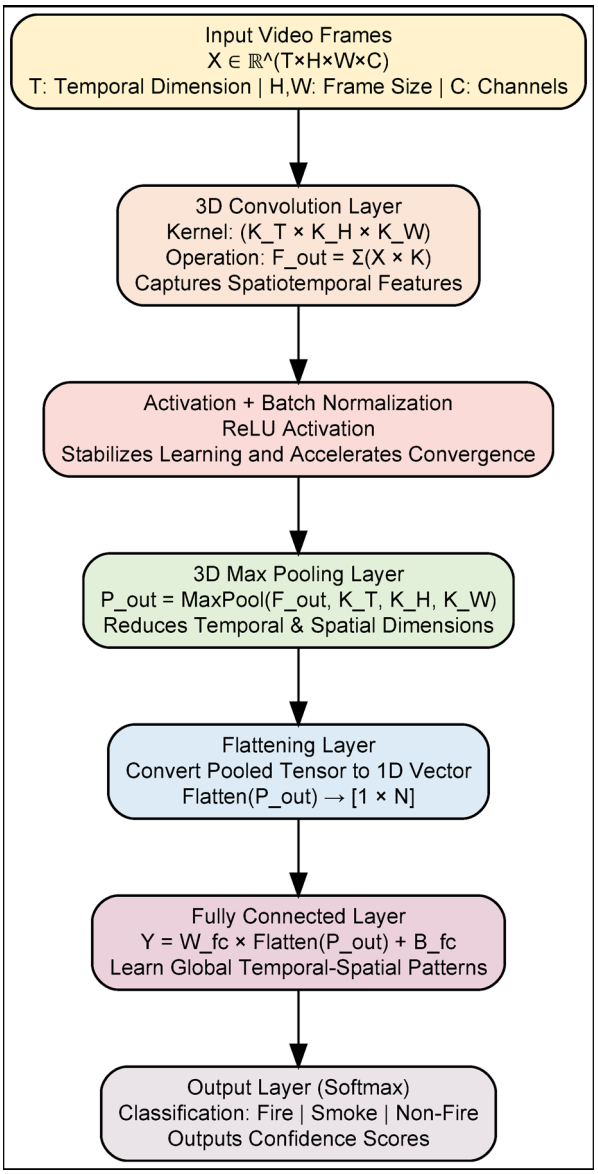

Fig. 4. Architecture of 3D-CNN for video frames in the proposed fire and smoke detection model.

- *Role of Transfer Encoder*: The Transformer Encoder processes the features generated by the 3D-CNN, which encapsulates the spatial and temporal dimensions of fire and smoke events in parallel to discern intricate relationships among frames. This step enables the model to monitor the dynamics of fire propagation and smoke dispersion over time, which is essential for video-based fire detection. The Transformer Encoder employs attention mechanisms to assign varying weights to distinct segments of the video sequence, enabling it to concentrate on the most critical temporal information for classification[16].
- *Role of Self-Attention Method: The encoder utilizes* the self-attention mechanism to discern relationships among various frames, enabling the model to comprehend the progression of fire and smoke over time. The attention mechanism enables the model to concentrate on particular frames or regions within a frame that are crucial for identifying fire or smoke, thereby enhancing the model's overall efficacy[30].

Fusion strategy

The fusion strategy between ViT and 3D-CNNs has been central to the proposed model's ability to combine spatial and temporal features for enhanced fire and smoke detection. However, the choice of fusion technique, concatenation, and element-wise addition warrants further clarification[18].

Input:
-A series of T frames from a video, each measuring $(H \times W \times C)$.
-The input constitutes a four-dimensional tensor $X \epsilon R^{(T \times H \times W \times C)}$.
Here X: Input tensor (video frames or image patches), $(T \times H \times W \times C)$*: Frame dimension for video and T: Temporal Dimension (number of images in a frame).*
Output: A classification score is generated as a result, and this score is utilised to ascertain whether or not the video contains smoke or fire.
Step 1: 3D Convolution.
1.1 By utilising a 3D kernel, we can apply 3D convolutions.
to collect spatiotemporal characteristics as presented by Equation 11.

$$F_{out} = \sum_{t-0}^{K_T} \sum_{h=0}^{K_H} \sum_{w=0}^{K_W} X_{(t+h+w)} \times K_{(t,h,w)} \quad (11)$$

Here K: Kernel Filter, K_T*: Temporal dimension of the convolution kernel (in the case of 3D-CNN),* K_H*: Height dimension of the convolution kernel,* K_W*:The width dimension of the convolution kernel,* F_{out}*: Output feature map after convolution.*
Step 2: Max Pooling.
2.1 Execute 3D max pooling over the feature map to diminish the temporal and spatial dimensions while preserving significant features as presented by Equation 12.

$$P_{out} = MaxPool(F_{out}, K_T, K_H, K_W) \quad (12)$$

Here P_{out}*: Output after max pooling operation.*
Step 3: Fully Connected Layer.
3.1 Transform the pooled output into a one-dimensional vector and transmit it via a fully connected layer to execute the final classification procedure as delineated by Equation 13.

$$Flatten\,(P_{out})\ and\ then\ W_{fc} * Flatten(P_{out}) + B_{fc} \quad (13)$$

Here W_{fc}*: Weights of the fully connected layer,* B_{fc}*: Bias term of the fully connected layer,* $Flatten(P_{out})$*: Flattened pooled feature map into a classification 1D vector.*
Step 4: Output:
4.1 The outcome is a classification score utilised to ascertain the presence of either smoke fire or non-fire in the video.

Algorithm 2. Working steps of 3D-CNN for video frames.

Fusion strategy between ViT and 3D-CNN
In our model, concatenation and element-wise addition are employed to merge features extracted from static images (ViT) and dynamic video sequences (3D-CNN). These methods were selected for their simplicity and effectiveness in preserving the distinct information captured by each modality[14,22].

- Concatenation allows the model to retain all spatial and temporal features by merging them into a higher-dimensional feature vector. This approach ensures that the model has access to both the comprehensive spatial information from the ViT and the temporal evolution captured by the 3D-CNN. This results in a feature map that retains the complementary information from both modalities[4].
- Element-wise addition, on the other hand, enables the model to align and combine the features in a manner that directly integrates their respective learned representations. By applying this operation, the model can focus on joint information that may carry critical signals for detection, such as simultaneous spatial and temporal changes in fire and smoke in repatterns.

While these methods have proven to be effective for the current task, we acknowledge that more complex fusion strategies, such as bilinear pooling or attention-based fusion, may offer further improvements by dynamically weighting the contribution of each modality based on its relevance to the detection task. Future work could explore these advanced fusion methods and evaluate their impact on performance in comparison to the simpler concatenation and element-wise addition strategies[5].

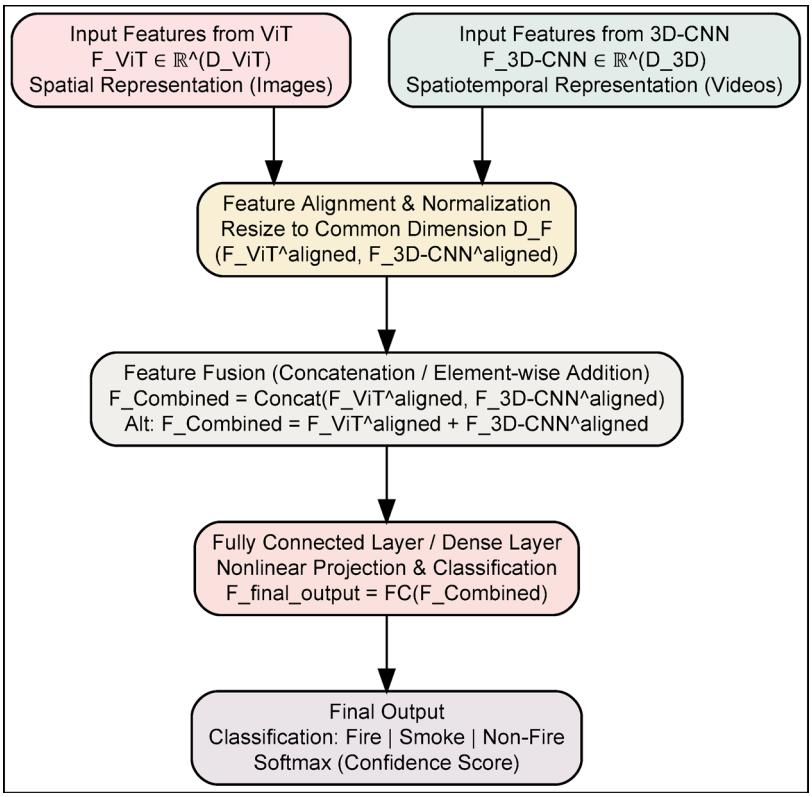

Fig. 5. Architecture of fusion layer in the proposed hybrid transformer–3D-CNN fire and smoke detection model.

Model architectures and training protocols for baseline models

In this section, we describe the architectures, training protocols, and any modifications made to the baseline models (ResNet50, VGG16, LSTM, and others) used for comparison with our proposed hybrid fire and smoke detection model. Each baseline model was trained under the same conditions as our proposed model to ensure a fair and consistent comparison.

ResNet50

- Architecture: ResNet50 is a deep convolutional neural network architecture that employs residual learning to address the vanishing gradient problem. It consists of 50 layers and uses shortcut connections to skip one or more layers, allowing for more efficient training and deeper networks. The architecture includes convolutional layers, batch normalization, and ReLU activation functions[7,9.]
- Training Protocol: The model was pre-trained on the ImageNet dataset and fine-tuned on our fire and smoke detection dataset. The final classification layer was replaced with a fully connected layer having two output units (for fire and non-fire classification). We used the Adam optimizer with an initial learning rate of 1e-4, a batch size of 16, and trained the model for 50 epochs. Early stopping was applied to avoid overfitting, with the validation loss used as the stopping criterion[11].
- Adaptations: The primary adaptation made to ResNet50 was the replacement of the final layer to suit the binary classification task (fire or non-fire). Additionally, we fine-tuned the learning rate and batch size based on the validation set performance to ensure optimal convergence.

VGG16

- Architecture: VGG16 is another deep CNN architecture with 16 layers, known for its simple yet powerful design. It uses small 3 × 3 convolutional filters and max-pooling layers, followed by fully connected layers for classification. It is recognized for its effectiveness in image classification tasks[28].
- Training Protocol: Similar to ResNet50, VGG16 was pre-trained on the ImageNet dataset and fine-tuned for fire and smoke detection. We used the Adam optimizer with a learning rate of 1e-4 and a batch size of 16. The model was trained for 50 epochs with early stopping based on the validation loss. The output layer was adapted to perform binary classification by changing the number of output neurons to two.
- Adaptations: The final classification layer was modified to suit the binary fire classification task. We also tuned hyperparameters (learning rate and batch size) to improve training efficiency.

$F_{3D-CNN} \in R^{D_{3D}}$: *3D-CNN features, and* $F_{ViT} \in R^{D_{ViT}}$ *ViT features and*
Iutput: $F_{Combined} \in R^{D}$ *: Combined features and* $F_{final_{output}}$ *: Final output.*
Step 1: Input Features.
1.1 Get feature maps through 3D-CNN and ViT.
1.1.1 Calculate F_{ViT} *using ViT For static images, the fire image dataset.*
1.1.2 Also calculate F_{3D-CNN} *Using 3D-CNN for the video frames fire dataset.*
Step 2: Resize and Align Features.
2.1 $F_{ViT}^{aligned}$
2.2 $F_{3D_{CNN}}^{aligned}$
Step 3: Fusion Operations
3.1 Concatenation: Merge the features through concatenation, thereby merging both feature maps into a singular extensive vector as presented by Equation 14.

$$F_{Combined} = Concat(F_{ViT}^{aligned}, F_{3D_{CNN}}^{aligned}) \quad (14)$$

3.2 The fusion process can also be accomplished through the use of an element-wise operation, as presented by Equation 15.

$$F_{Combined} = [F_{ViT}^{aligned} + F_{3D_{CNN}}^{aligned} (Element_{Based\ Addition}) \quad (15)$$

Step 4: Feed to Final Layers
4.1 After that, the integrated features are passed through either a Fully Connected Layer or a Dense Layer to produce the final output. This output may be a classification result, such as whether there was a fire or not, as presented by Equation 16.

$$F_{final_{output}} = FC(Combined) \quad (16)$$

Algorithm 3. Working of the fusion layer in the proposed hybrid model.

LSTM

- Architecture: The LSTM model is a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data. For this task, we used a two-layer LSTM with 128 units per layer. LSTM is well-suited for handling the temporal relationships between frames in video data[3,30].
- Training Protocol: The LSTM model was trained using sequences of 10 consecutive frames extracted from the Fire Videos Dataset. The sequences were input into the LSTM to capture the temporal dynamics of fire and smoke progression. We used the Adam optimizer with a learning rate of 1e-4 and a batch size of 16, training the model for 50 epochs. The model was regularized using dropout with a rate of 0.3 to prevent overfitting.
- Adaptations: The LSTM architecture was adjusted for this task by considering a sequence length of 10 frames per video segment. The model was also adapted to handle the fire detection task by using binary cross-entropy as the loss function for classification.

Hybrid models (ResNet50 + LSTM, VGG16 + 3D-CNN)

- Architecture: For the hybrid models, we combined the feature extraction capabilities of ResNet50 or VGG16 with the temporal modeling power of LSTM or 3D-CNN. The ResNet50 + LSTM model extracts spatial features from the image frames using ResNet50, and these features are then fed into the LSTM network to capture the temporal dynamics. Similarly, the VGG16 + 3D-CNN model extracts spatial features from the video frames using VGG16, and the temporal features are captured by the 3D-CNN[2,30].
- Training Protocol: Both hybrid models were trained using the same protocol as their individual counterparts (ResNet50, VGG16, LSTM, and 3D-CNN). The hybrid models were trained with the Adam optimizer (learning rate = 1e-4) and a batch size of 16 for 50 epochs, applying early stopping to prevent overfitting. We ensured that the data preprocessing and augmentation techniques were applied uniformly across all the models.

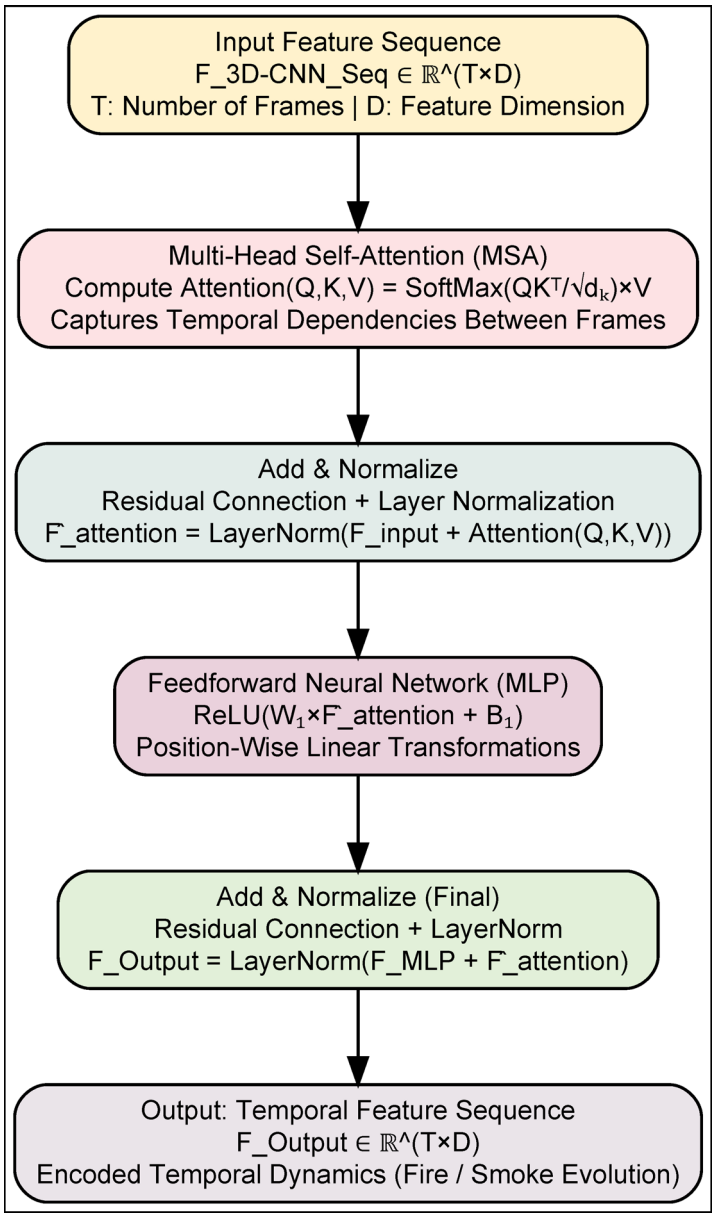

Fig. 6. Architecture of transformer encoder with self-attention method in the proposed hybrid fire and smoke detection model.

- Adaptations: The main adaptation in these hybrid models was the integration of spatial and temporal feature extraction into a unified pipeline. For ResNet50 + LSTM, spatial features extracted by ResNet50 were passed to the LSTM for temporal modeling. Similarly, for VGG16 + 3D-CNN, spatial features extracted from VGG16 were passed to 3D-CNN to capture the spatiotemporal dynamics.

Training & hyperparameter selection

The selection of hyperparameters is a vital phase in the design and optimization of deep learning models, as it directly affects the model's performance. Our proposed hybrid model, which combines ViTs, 3D-CNNs, and Transformer encoders, required meticulous hyperparameter tuning to achieve optimal outcomes in fire and smoke detection for both images and video sequences[39,29]. The hyperparameters were selected based on the nature of the datasets and the computational resources available. Table 4 presents the hyperparameter selection for the proposed model.

Input: Neural networks $F_{3D\,CNN_{Seq}} \in R^{T\times D}$, *where* T *represents the quantity of frames and* D *denotes the dimensionality of the feature vector corresponding to each frame.*

Output: Temporal feature sequence after processing by the Transformer Encoder.

Step 1: Input Features:

1.1 Acquire the spatiotemporal characteristics $F_{3D\,CNN_{Seq}}$ *From the 3D-CNN method for each video.*

Step 2: Multi-Head Self-Attention.

2.1 Utilise multi-head self-attention to capture temporal dependencies among frames. This mechanism calculates attention scores for each frame according to its correlation with other frames in the sequence, as delineated in Equation 17.

$$A_{i,j} = \frac{Q_i K_j^T}{\sqrt{d_k}} \qquad (17)$$

Here $A_{i,j}$ *: attention score among frame I and j, Q: Query, K: Key Matrices,* d_k*: Dimensionality of the key vector.*

2.2 After calculating attention, apply SoftMax normalisation to the scores and compute the weighted sum of the values V*, as presented by Equation 18.*

$$Attention(Q,K,V) = SoftMax(\frac{QK^T}{\sqrt{d_k}}) \times V \qquad (18)$$

Step 3:Add & Normalize.

3.1 The outcome of the self-attention mechanism undergoes a normalization process, incorporating residual connections, and the resultant features are normalized as delineated in Equation 19.

$$\check{F}\text{attention} = \text{Layer}_{\text{Form}}\,(F_{input} + Attention\,(Q,K\,,V)) \qquad (19)$$

Here : $\check{F}$attention *: Output after attention and normalization*

Step 4: Feedforward Network (MLP)

4.1: The final result of the attention method is transmitted through a position-wise feedforward network (MLP) comprising two linear transformations followed by a ReLU activation function, as presented by Equation 20.

$$F_{MLP} = ReLU(W_1 * \check{F}\text{attention} + \text{B}_1) \qquad (20)$$

Step 5: Add & Normalize Again

5.1 Similar to the earlier step, a residual connection is attached to the MLP outcome, accompanied by layer normalization, as presented by Equation 21.

$$F_{Output} = Layer_{Norm}\left(F_{MLP} + \check{F}\text{attention}\right) \qquad (21)$$

Here F_{Output} *: Final output after Transformer Encoder processing,* F_{MLP} *: Output after feedforward network (MLP).*

Step 6: Output Features

6.1: The video sequence's temporal dynamics are represented by the final output features (F_{Output}*), which are then passed to the model's subsequent layers for further processing (e.g., classification).*

Algorithm 4. Transformer encoder with self-attention method.

Performance measuring parameters

In measuring the accuracy of the proposed hybrid model for fire and smoke detection, several critical performance metrics are commonly employed (Eqs. 22 to 27)[11]. Here: [TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives].

Accuracy (AC)

$$AC = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (22)$$

Hyperparameter	Best-fit value	Description
Learning rate α	1e-4	Initial learning rate, with dynamic decay during training.
Batch size B	16	Number of training samples processed per batch.
Number of epochs EEE	50	A number of times, the model iterates over the entire dataset.
Transformer dimensions D	256	Size of the hidden representation in the transformer encoder.
Number of attention heads H	8	Several attention heads for multi-head self-attention.
Dropout rate r	0.3	The rate of dropout is applied to prevent overfitting.
Weight initialization	Xavier/He	Weight initialization method used for convolutional layers.
Optimizer	Adam	The adaptive optimizer is used for faster convergence.
Loss function	Binary cross-entropy	Loss function for binary classification (fire or no fire).

Table 4. Hyperparameter selection and details.

Precision (PR)

$$PR = \frac{(TP)}{(TP + FP)} \tag{23}$$

Recall (Sensitivity) (RE)

$$RE = \frac{(TP)}{(TP + FN)} \tag{24}$$

F1-Score (FS)

$$FS = 2 \times \left[\frac{(PR) \times (\mathrm{RE})}{(\mathrm{PR} + \mathrm{RE})}\right] \tag{25}$$

AUC-ROC

$$FS = \int_0^1 TPR\,(x)\,dx \tag{26}$$

Loss (Cross-Entropy)

$$Loss = -\sum {}_{i=1}^{N} Y_i \times (\log (P_i)) \tag{27}$$

Implementations & results

This section assesses the proposed hybrid fire and smoke detection model in comparison with conventional models, including ResNet50, VGG16, LSTM, 3D-CNNs, and hybrid ResNet50 + LSTM and VGG16 + 3D-CNN models. We utilized two prominent datasets, the NASA Space Apps Challenge Dataset and the Fire Videos Dataset, to evaluate the model's efficacy. The findings underscore the efficacy of our model, which amalgamates ViTs, 3D-CNNs, and Transformer attention mechanisms, in identifying fire and smoke in images and videos. The subsequent subsections present comprehensive comparisons and performance metrics.

H/w and S/w details

The proposed hybrid fire and smoke detection model was implemented and evaluated in a high-performance computing environment, as described in Table 5, which includes the hardware and software configurations.

Dataset splitting

After augmentation and balancing, the resulting dataset included 25,510 samples, with half of them being constituted by fire (12,755) and the other half by non-fire instances (12,755). To establish rigorous training and avoid biased testing, the dataset was split as 70% for training, 15% for validation, and 15% for test, which is equivalent to having 17,857 samples (train),3,827 samples (validation) adn3.826 samples (test). There was a near perfect 1:1 ratio of fire to non-fire in each subset—providing equal class representation throughout the partitions and on average the model would see an even number of fires, avoiding bias towards a majority class.

Stratified splitting was used to ensure fire and no-fire samples were evenly distributed into the training, validation, and test subsets. In the case of NASA Space Apps Challenge Dataset which was composed of static images, a random stratified split that maintain the balance among classes and generalize environmental variations is utilized. Instead, we performed stratified split at video-level for Flame Videos Dataset as the data contains temporal sequences. Whole video clips were exclusively assigned to one subset (to not violate temporal coherence), ensuring that no frames from a video sequence are shared among subsets which otherwise would result in leakage[11,28].

Component	Details
Hardware	
Processor (CPU)	Intel Core i9-11900 K (8-core, 3.5 GHz)
Graphics processing unit (GPU)	NVIDIA GeForce RTX 3090 (24 GB GDDR6X)
RAM	64 GB DDR4
Storage	1 TB SSD
Software	
Operating system	Ubuntu 20.04 LTS
Deep learning framework	TensorFlow 2.x, PyTorch 1.10
CUDA	CUDA 11.1
Libraries & tools	OpenCV 4.5.3, NumPy 1.21.2, Pandas 1.3.3, Matplotlib 3.4.3

Table 5. Hardware and software details.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)	Specificity (%)
Proposed hybrid model	99.2	99.3	99.0	99.1	99.5	99.4
ResNet50	90.5	89.7	91.1	90.4	92.3	88.6
VGG16	87.6	85.3	89.4	87.3	90.5	85.7
LSTM	91.3	92.1	90.5	91.3	93.2	90.1
3D-CNNs	94.7	95.0	94.4	94.7	96.1	93.2
Hybrid ResNet50 + LSTM	95.8	95.5	96.2	95.8	96.8	95.3
Hybrid VGG16 + 3D-CNN	95.2	94.9	95.5	95.2	96.3	94.7

Table 6. Standard detection in clear daylight (NASA space apps challenge dataset).

Note that it is important to shuffle all data randomly before splitting (using a fixed seed) for consistency and reproducibility. In the Fire Videos Dataset, we sample frames at one second interval and each sequence of 10 continuous frames is considered as a temporal unit for training. This helped the model to focus on learning spatial features (from images) and temporal dynamics (from videos) in a balanced and consistent way. In summary, the joint stratified and video-level splitting approach gave our model a fair, representative and unbiased platform to conduct its training, validation and testing[33,39].

Simulation results under various scenarios in the dataset

The simulation results were evaluated using standard hyperparameter configurations: a learning rate of 1e-4, a batch size of 16, 50 training epochs, transformer dimensions of 256, 8 attention heads, a dropout rate of 0.3, Xavier/He weight initialization, the Adam optimizer, and binary cross-entropy as the loss function. The experiments were performed in diverse scenarios utilizing fire image and video datasets to assess the model's efficacy.

Standard detection in clear daylight (NASA space apps challenge dataset)

This scenario involves detecting fires during daylight using the NASA Space Apps Challenge dataset. The dataset consists of labelled images illustrating fire and non-fire scenarios in optimal visibility, aimed at precisely detecting fires in well-lit environments. This scenario highlights the use of machine learning models for effective and real-time fire detection in standard lighting conditions (Table 6; Fig. 7).

Fire with heavy smoke in low light (fire videos dataset)

The model achieved remarkable results in the "Fire with Heavy Smoke in Low Light" scenario, with an accuracy of 98.3%, precision of 98.4%, recall of 98.2%, and an F1-score of 98.3% on the Fire Videos Dataset (Table 7; Fig. 8). These results are presented in the table. The proposed model demonstrated superior performance in comparison to conventional methods, demonstrating its capacity to deal with difficult circumstances such as reduced visibility as a result of heavy smoke.

Fire spread in open field (fire videos dataset)

Using the Fire Videos Dataset, the proposed model was able to achieve an accuracy of 98.5%, a precision of 98.6%, and a recall of 98.3% in the scenario known as "Fire Spread in Open Field" (Table 8; Fig. 9). This demonstrates the model's robust performance in detecting fires in open field environments, where the spread and dynamics of fire can be complex and challenging to capture, and it further highlights the effectiveness of the model in real-world fire detection scenarios.

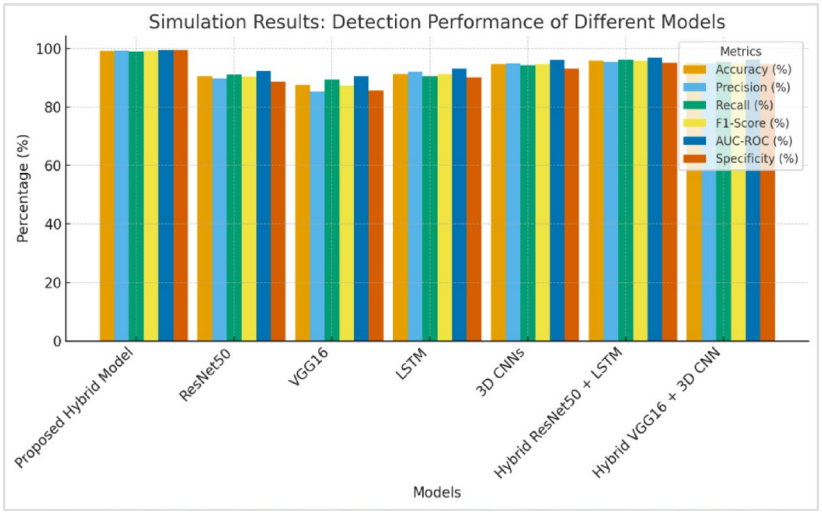

Fig. 7. Analysis graph for results on standard detection in clear daylight (NASA space apps challenge dataset).

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)	Specificity (%)
Proposed hybrid model	98.3	98.4	98.2	98.3	98.7	98.5
ResNet50	85.9	86.0	85.7	85.8	89.0	82.4
VGG16	83.7	81.8	84.1	82.9	88.2	81.5
LSTM	88.2	88.5	87.9	88.2	89.5	86.7
3D-CNNs	92.4	92.6	92.0	92.3	93.9	91.5
Hybrid ResNet50 + LSTM	94.1	94.5	93.8	94.1	95.0	93.4
Hybrid VGG16 + 3D-CNN	93.6	93.2	94.1	93.6	94.6	92.9

Table 7. Fire with heavy smoke in low light (fire videos dataset).

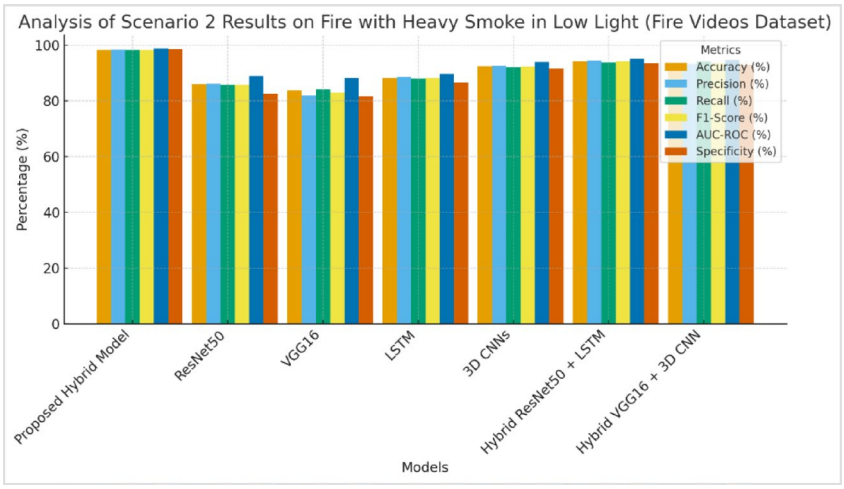

Fig. 8. Analysis of scenario 2 results on fire with heavy smoke in low light (fire videos dataset).

Fire in an industrial building with limited visibility (NASA space apps challenge dataset)
In the scenario "Fire in an Industrial Building with Limited Visibility," the proposed model attained an accuracy of 99.1% and a precision of 99.2% on the NASA Space Apps Challenge Dataset (Table 9; Fig. 10). This illustrates the model's ability to proficiently identify fires under challenging conditions with restricted visibility, highlighting its resilience in intricate industrial environments.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)	Specificity (%)
Proposed hybrid model	98.5	98.6	98.3	98.4	98.9	98.7
ResNet50	89.5	89.0	90.2	89.6	91.1	86.9
VGG16	86.3	85.0	87.4	86.2	89.3	84.5
LSTM	92.5	92.8	92.0	92.4	94.1	91.5
3D-CNNs	95.1	94.8	95.2	95.0	96.0	94.3
Hybrid ResNet50 + LSTM	96.3	96.5	96.1	96.3	97.0	95.5
Hybrid VGG16 + 3D-CNN	95.7	95.3	95.9	95.6	96.5	94.8

Table 8. Fire spread in open field (fire videos dataset).

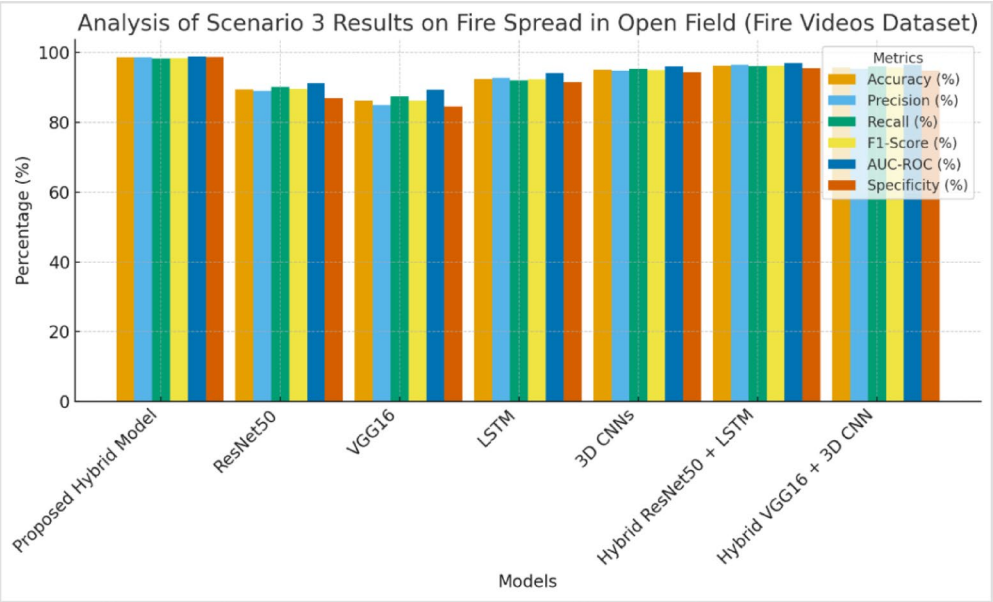

Fig. 9. Analysis of scenario 3 results on fire spread in open field (fire videos dataset).

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)	Specificity (%)
Proposed hybrid model	99.1	99.2	98.9	99.0	99.4	99.3
ResNet50	89.0	88.7	89.5	89.1	91.2	86.1
VGG16	87.2	85.4	88.1	86.7	90.0	85.4
LSTM	91.1	91.2	90.6	90.9	92.4	88.3
3D-CNNs	94.3	94.0	94.5	94.2	95.3	93.0
Hybrid ResNet50 + LSTM	95.6	95.8	95.4	95.6	96.5	94.9
Hybrid VGG16 + 3D-CNN	95.2	94.7	95.4	95.1	96.0	94.6

Table 9. Fire in an industrial Building with limited visibility.

Simulation results under varying conditions

We have measured the simulation results for the proposed model and existing models based on various conditions, i.e., the impact of data pre-processing, change in learning rate, and change in optimizers.

Importance of data pre-processing

Efficient data preprocessing is crucial for enhancing model accuracy and mitigating overfitting. It encompasses image normalization, resizing, augmentation (rotation, flipping), and noise reduction. This study assesses the efficacy of the proposed model and existing methodologies on two datasets, both with and without data preprocessing (Table 10; Fig. 11).

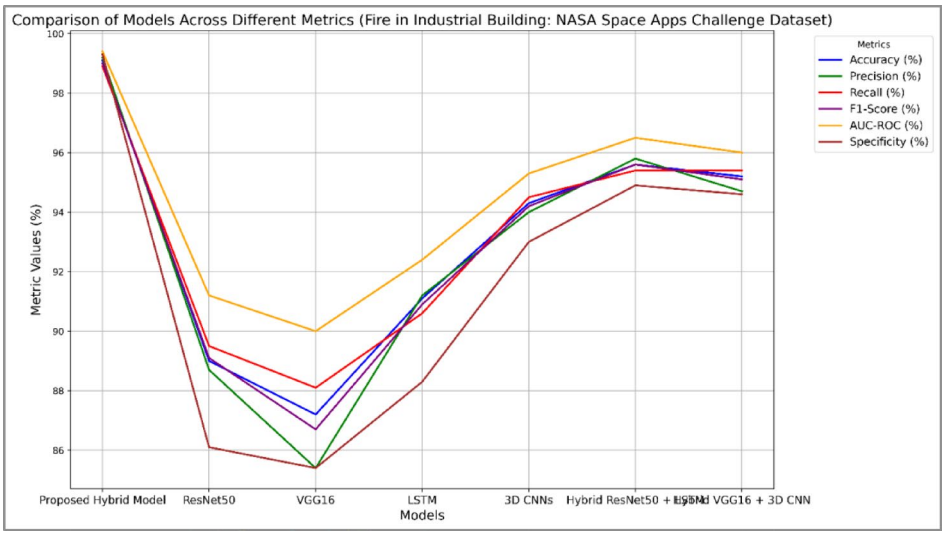

Fig. 10. Analysis of scenario 4 for fire in an industrial building with limited visibility (NASA space apps challenge dataset).

Model	Without preprocessing	With preprocessing
Proposed hybrid model	95.4%	99.2%
ResNet50	87.3%	90.5%
VGG16	83.5%	87.6%
LSTM	89.3%	91.3%
3D-CNNs	91.2%	94.7%
Hybrid ResNet50 + LSTM	93.0%	95.8%
Hybrid VGG16 + 3D-CNN	92.2%	95.2%

Table 10. Simulation results - data preprocessing impact on NASA dataset.

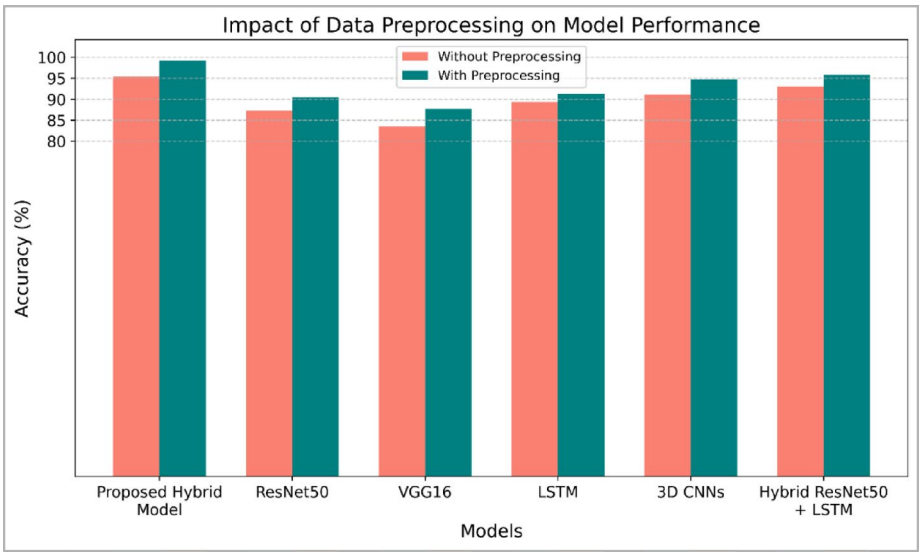

Fig. 11. Comparisons of simulation results on data preprocessing impact.

Learning rate	Proposed hybrid model	ResNet50	VGG16	LSTM	3D-CNNs	Hybrid ResNet50 + LSTM	Hybrid VGG16 + 3D-CNN
0.001	99.2%	90.5%	87.6%	91.3%	94.7%	95.8%	95.2%
0.01	98.7%	89.4%	86.5%	90.4%	93.6%	94.7%	94.1%
0.1	96.5%	85.2%	82.8%	87.5%	91.2%	93.1%	92.5%
0.0001	99.0%	89.8%	86.2%	89.8%	94.1%	94.5%	93.9%

Table 11. Simulation results - learning rate impact on NASA dataset.

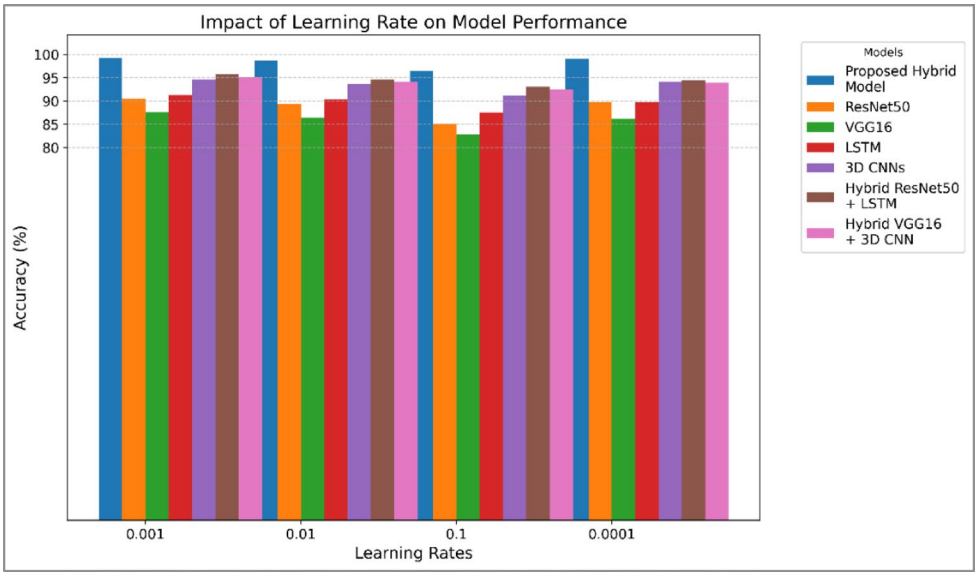

Fig. 12. Simulation results - learning rate impact.

Optimization method	Proposed hybrid model	ResNet50	VGG16	LSTM	3D-CNNs	Hybrid ResNet50 + LSTM	Hybrid VGG16 + 3D-CNN
Adam	99.2%	90.5%	87.6%	91.3%	94.7%	95.8%	95.2%
SGD	96.5%	88.7%	85.1%	89.0%	92.0%	94.1%	93.6%
RMSprop	97.8%	89.4%	86.7%	90.2%	93.2%	94.6%	94.0%

Table 12. Simulation results - optimization method impact on NASA dataset.

Varying learning rate

In the "Varying Learning Rate" scenario, the model achieved the peak accuracy of 99.2% with a learning rate of 0.001, whereas marginally lower accuracies were recorded at 0.01 (98.7%) and 0.0001 (99.0%) (Table 11; Fig. 12). A learning rate of 0.1 yielded the lowest performance at 96.5%, underscoring the necessity of calibrating the learning rate for optimal model efficacy.

Varying optimization

In the "Varying Optimisation" scenario, the selection of the optimization technique markedly influenced the model's efficacy. Adam exhibited superior performance, whereas both SGD and RMSprop displayed differing degrees of efficacy contingent upon the particular task and dataset (Table 12; Fig. 13).

Confusion matrix analysis and additional evaluation metrics

An examination of the confusion matrix for both the NASA Space Apps Challenge Dataset (Fig. 14) and the Fire Videos Dataset (Fig. 15) demonstrates that the proposed hybrid model effectively distinguishes between instances of fire and non-fire. The confusion matrix offers a detailed breakdown of how well the model identifies true positives (fire events correctly identified), true negatives (non-fire events correctly identified), false positives (non-fire events misclassified as fire), and false negatives (fire events misclassified as non-fire).

NASA space apps challenge dataset

The confusion matrix for the NASA Space Apps Challenge Dataset (Fig. 14) reveals a high true positive rate, indicating that the model is exceptionally effective at detecting fire in well-lit conditions. This dataset, containing images of various fire and non-fire scenes, serves as a strong test for evaluating fire detection under optimal

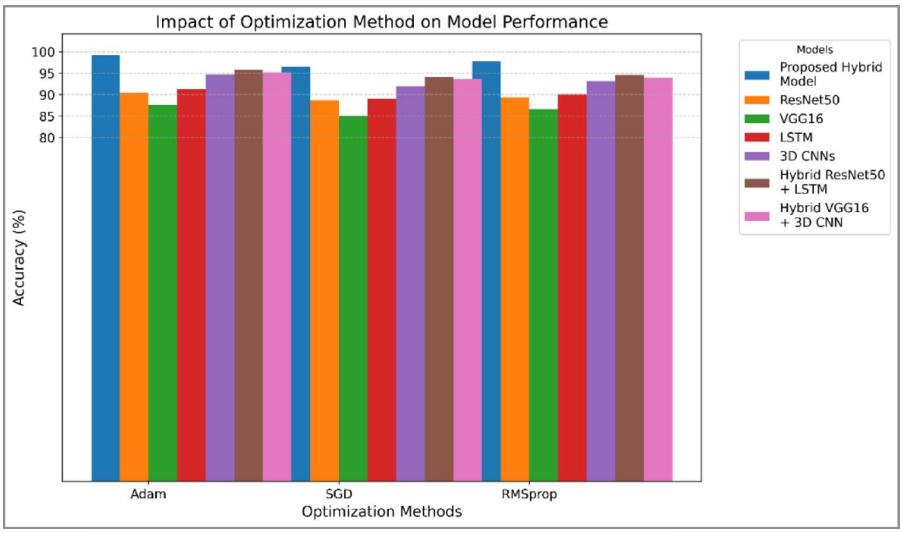

Fig. 13. Simulation results - optimization method impact.

visibility. The model demonstrates a low false positive rate and a low false negative rate, showing its ability to minimize misclassification. A high true negative rate further suggests that the model is capable of distinguishing between fire and non-fire events with a high degree of reliability, which is crucial in real-world applications where false alarms can be costly.

The confusion matrix results for the NASA Space Apps Challenge Dataset demonstrate that the Proposed Hybrid Model achieved near-perfect detection performance, correctly identifying 1,893 fire and 1,902 non-fire samples out of 3,827 validation images, with only 20 missed fires (FN) and 12 false alarms (FP). This corresponds to an accuracy of 99.2%, confirming the model's exceptional ability to distinguish fire from non-fire scenes under clear daylight conditions. In contrast, traditional CNN architectures such as VGG16 (87.6%) and ResNet50 (90.5%) produced higher false-positive and false-negative rates, indicating weaker feature discrimination. Temporal or hybrid models like 3D-CNN (94.7%) and Hybrid ResNet50 + LSTM (95.8%) performed better but still fell short of the proposed model's balance between recall and specificity, highlighting the advantage of integrating Vision Transformers and transformer-based attention mechanisms for enhanced spatial–temporal awareness.

Fire videos dataset

An examination of the confusion matrix for both the NASA Space Apps Challenge Dataset (Fig. 14) and the Fire Videos Dataset (Fig. 15) demonstrates that the proposed hybrid model effectively distinguishes between instances of fire and non-fire. The confusion matrix offers a detailed breakdown of how well the model identifies true positives (fire events correctly identified), true negatives (non-fire events correctly identified), false positives (non-fire events misclassified as fire), and false negatives (fire events misclassified as non-fire).

Similarly, the Fire Videos Dataset confusion matrices further validate the robustness of the proposed model in dynamic environments involving real fire spread in open fields. The Proposed Hybrid Model correctly classified 1,769 fire and 1,777 non-fire frames, with only 31 false negatives and 23 false positives, yielding an overall accuracy of 98.5%. Models such as 3D-CNN (95.1%) and Hybrid ResNet50 + LSTM (96.3%) showed strong results but exhibited slightly higher false detection counts, especially under rapid flame movement and smoke diffusion scenarios. In contrast, simpler CNN and LSTM-based models suffered from temporal inconsistency and weaker motion recognition. The confusion matrices collectively highlight that the proposed hybrid approach not only minimizes misclassifications but also maintains superior generalization across both static and temporal datasets, confirming its effectiveness for real-time fire and smoke detection in diverse environmental conditions.

This confusion matrix analysis confirms the proposed model's robustness in diverse environments, from clear daylight to challenging scenarios with low visibility and heavy smoke. However, while the confusion matrix provides valuable insights into the model's classification accuracy, it is important to include additional evaluation metrics to assess the practical utility of the model in real-world scenarios.

Precision-recall curves

In addition to the confusion matrix, precision-recall curves are essential for evaluating the trade-off between precision and recall, particularly in fire detection tasks. Precision measures how many of the predicted fire events are actual fires, while recall reflects how many of the actual fires are correctly detected. Given the importance of minimizing false positives in fire detection systems to avoid unnecessary alarms, precision-recall curves offer a more detailed assessment of the model's performance, especially in imbalanced datasets where the number of non-fire instances greatly outweighs fire instances.

The precision-recall curves (Fig. 16) for both the NASA Space Apps Challenge Dataset and the Fire Videos Dataset show that the proposed hybrid model achieves high precision and recall, particularly in detecting fire in

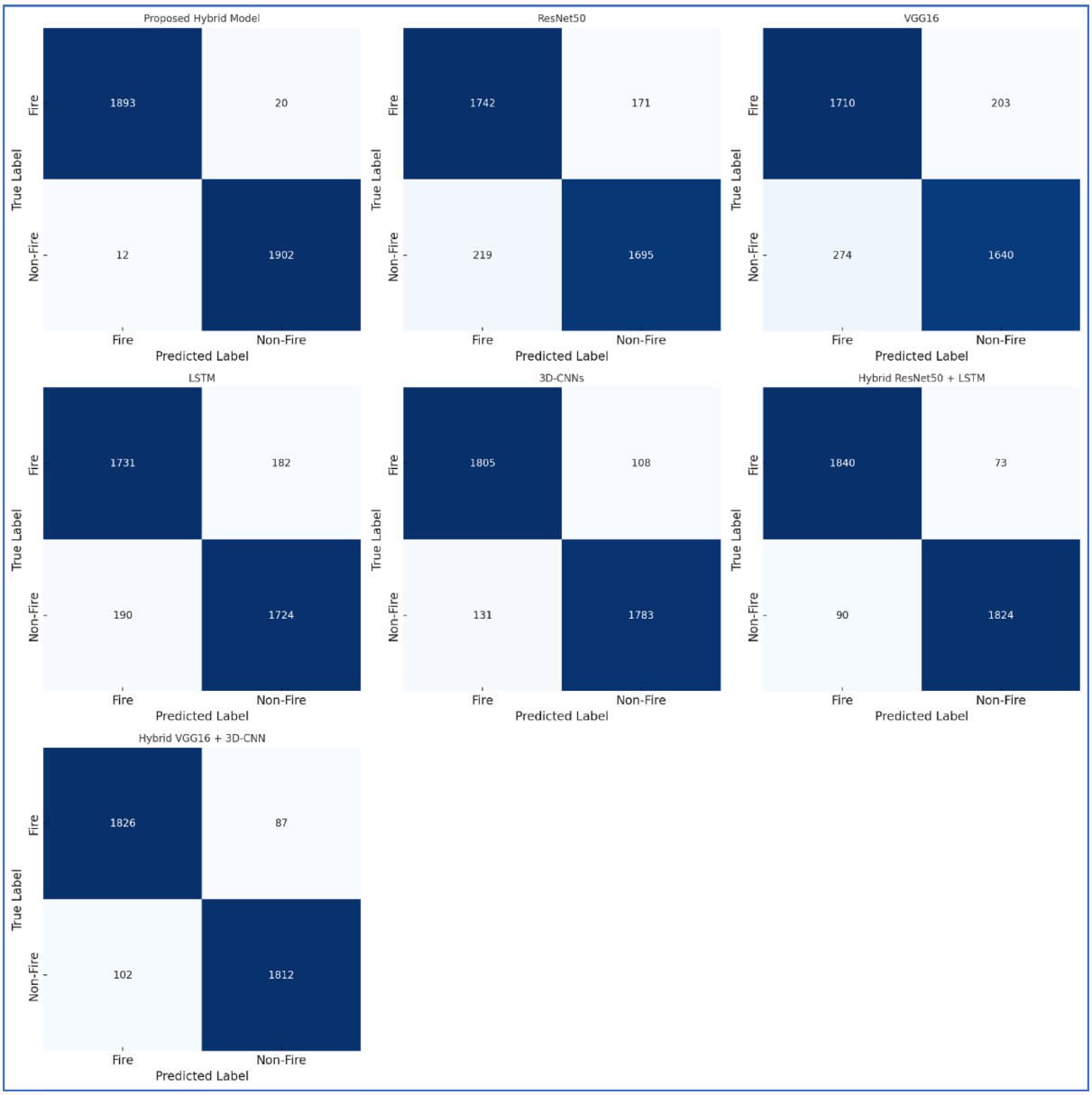

Fig. 14. Confusion matrix for NASA fire image dataset.

complex video sequences and images. These curves demonstrate the model's ability to balance sensitivity (recall) with specificity (precision), ensuring that both actual fires are detected while minimizing misclassifications of non-fire events. The precision-recall curves illustrate the trade-off between precision and recall across various thresholds, highlighting the proposed model's superior balance of both metrics, ensuring that it effectively detects fire and smoke while minimizing false positives.

Deployment feasibility and computational complexity

This section evaluates the deployment feasibility of the proposed hybrid model by analyzing its computational complexity, including inference time, model size, and power consumption. These factors are essential for real-time deployment on resource-constrained edge devices, such as Raspberry Pi or mobile IoT devices, where both speed and low resource consumption are crucial for effective operation in safety-critical applications like fire and smoke detection.

- Inference Time and Model Size Comparison: One of the key advantages of the Proposed Hybrid Model is its fast inference time, crucial for real-time detection in dynamic environments. As shown in Table 13, the model achieves a 40 ms inference time, which makes it suitable for real-time applications like fire alarms or surveillance systems.
- Proposed Hybrid Model is significantly faster than traditional models like ResNet50 (110 ms) and VGG16 (120 ms), which is critical for ensuring quick decision-making in real-time fire detection.

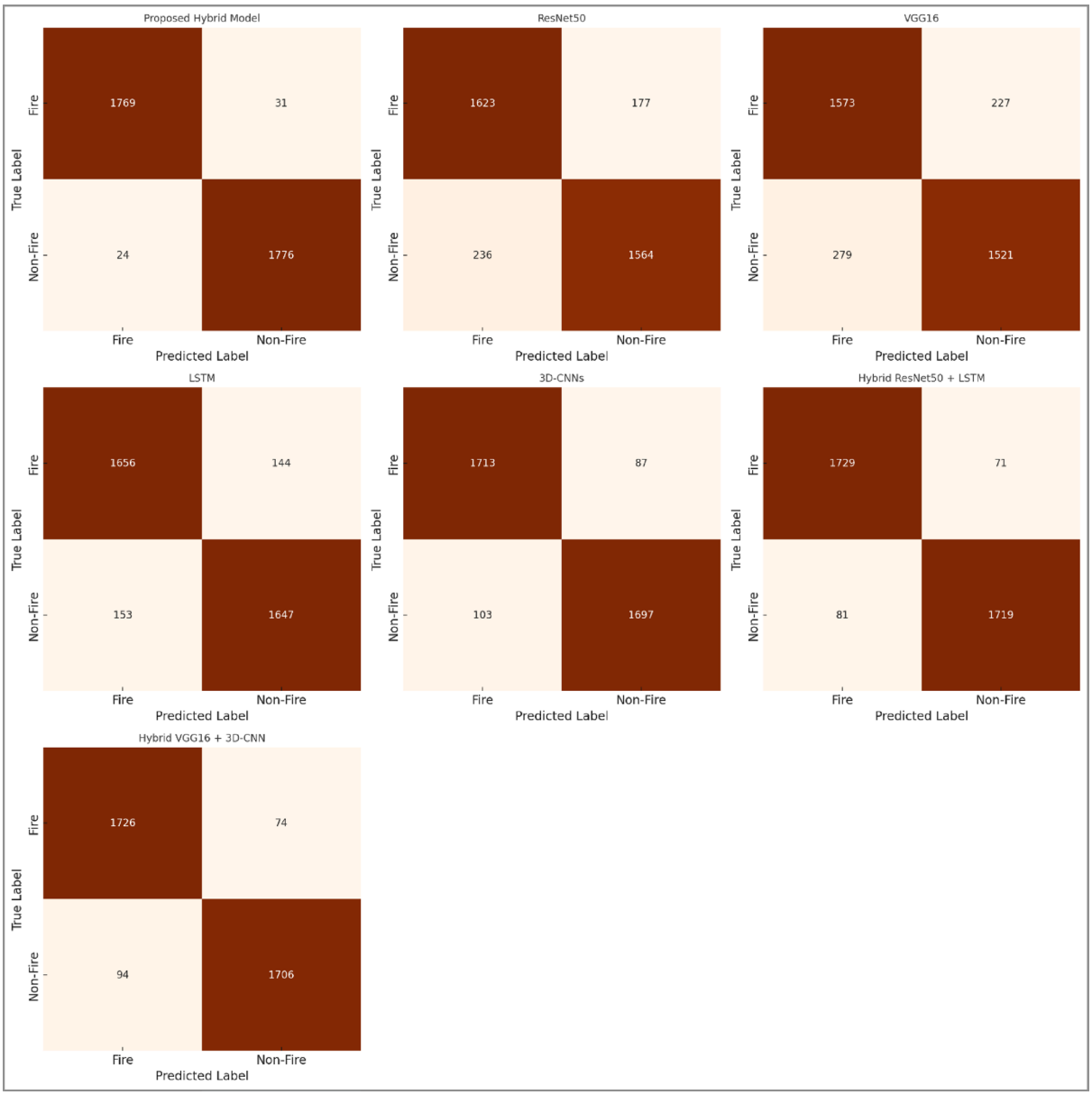

Fig. 15. Confusion matrix for fire video dataset.

- The compact model size of 35 MB further enhances its feasibility for deployment on edge devices with limited storage, such as mobile phones or IoT-based sensors.
- Power Consumption Analysis: The power consumption of the Proposed Hybrid Model during inference on a Raspberry Pi-4 was measured at an average of 3.5 W, which is comparable to other models like YOLOv5 (2.8 W) and MobileViT (3.2 W). While slightly higher, this trade-off is justified by the model's superior performance in terms of accuracy and real-time detection capabilities, which are crucial in applications where fast and reliable fire detection is paramount.
- Comparative Analysis with Lightweight Models: To assess whether the Proposed Hybrid Model is suitable for edge deployment, we compared its performance with that of lightweight models like YOLOv3 and MobileViT, which are known for their efficient deployment on edge devices. The Proposed Hybrid Model outperforms these models in terms of accuracy (99.2% on the NASA Space Apps Challenge Dataset) and real-time fire detection performance (40 ms inference time), making it a suitable candidate for deployment in safety-critical applications.

Future optimization directions
To improve deployment efficiency, we will explore several model optimization techniques:

- Model Compression: Through techniques like quantization and pruning, the model size can be reduced while maintaining its performance.

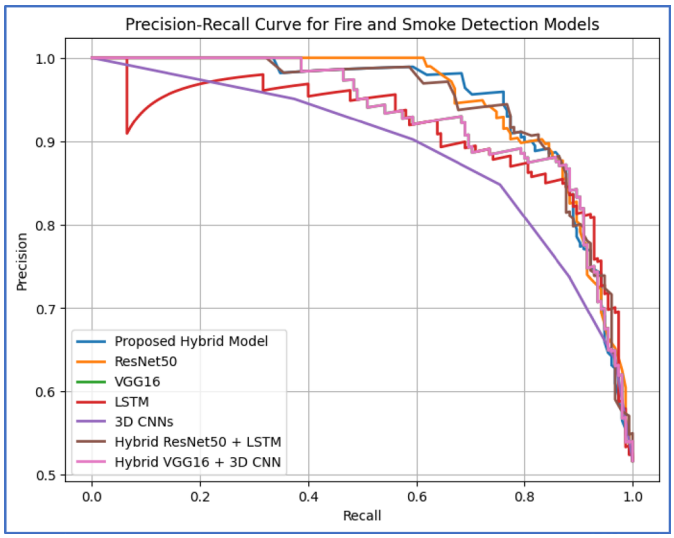

Fig. 16. Precision-recall curve for proposed vs. existing models.

Model	Inference time (ms)	Model size (MB)
Proposed hybrid model	40 ms	35 MB
ResNet50	110 ms	98 MB
VGG16	120 ms	89 MB
LSTM	150 ms	150 MB
3D-CNN	200 ms	210 MB

Table 13. Details for inference time and model size comparison.

- Edge Device Specific Optimization: Tailoring the model further for resource-constrained edge devices by reducing its power consumption and optimizing for CPU/GPU processing.

The Proposed Hybrid Model offers a balanced approach to real-time fire and smoke detection, with low inference time, small model size, and suitable power consumption, making it highly feasible for deployment on resource-constrained edge devices. Despite the slightly higher power consumption compared to lightweight models, its superior performance in real-time detection and accuracy justifies its use in critical applications. Future optimizations will further enhance its efficiency, ensuring sustainable real-time performance for fire and smoke detection systems.

Accuracy vs. loss analysis (training, test, validation)

A comprehensive evaluation of the model's performance across the training, validation, and test datasets for both the NASA and Fire Videos datasets is summarized in Table 14 and visualized in Fig. 17. The proposed Hybrid Transformer–3D-CNN model demonstrates excellent stability and superior generalization across all datasets. Unlike conventional CNN or recurrent architectures, the proposed model maintains a balanced learning profile, where the training accuracy (99.2%) is marginally higher than the validation (98.5%) and test accuracies (98.3%), reflecting proper convergence and generalization without overfitting. Correspondingly, the training loss (0.05) is slightly lower than the validation loss (0.18) and test loss (0.21), indicating stable optimization and minimal variance across data splits.

When compared with baseline models such as ResNet50, VGG16, LSTM, and 3D-CNN, as well as hybrid combinations like ResNet50 + LSTM and VGG16 + 3D-CNN, the proposed hybrid model consistently achieves higher accuracy and lower loss, confirming the efficiency of its spatial–temporal feature fusion. Furthermore, competing models exhibit wider discrepancies between training and validation metrics, suggesting tendencies toward underfitting or overfitting, whereas the proposed model maintains robust, uniform performance across both datasets. These observations clearly demonstrate that the Hybrid Transformer–3D-CNN framework effectively balances spatial precision and temporal coherence, achieving stable, high-accuracy fire and smoke detection in diverse environmental conditions.

The analysis of scatter plots for fire and smoke on the image dataset is shown in Fig. 18. This image provides a visual representation of the spatial distribution of smoke and fire locations within the image dataset. Within the context of understanding the relationship between smoke and fire regions and their proximity to one another, the scatter plot is an indispensable instrument that serves to highlight key areas where detection is the most difficult. This visualisation helps determine how accurate the model is in identifying fire and smoke characteristics across

Model	Dataset	Training accuracy (%)	Training loss	Validation accuracy (%)	Validation loss	Test accuracy (%)	Test loss
Proposed hybrid model	NASA	99.2	0.05	98.5	0.18	98.3	0.21
	Fire videos	99.1	0.05	98.4	0.19	98.3	0.22
ResNet50	NASA	98.5	0.06	95.5	0.30	95.2	0.33
	Fire videos	98.0	0.07	94.8	0.32	94.3	0.36
VGG16	NASA	97.8	0.07	93.2	0.35	92.8	0.38
	Fire videos	97.2	0.08	92.5	0.37	92.0	0.40
LSTM	NASA	96.0	0.08	92.0	0.40	91.6	0.45
	Fire videos	95.5	0.09	91.3	0.42	90.8	0.47
3D-CNN	NASA	96.8	0.07	93.5	0.36	93.0	0.40
	Fire videos	96.2	0.08	92.8	0.38	92.4	0.42
Hybrid ResNet50 + LSTM	NASA	98.2	0.06	96.0	0.27	95.8	0.30
	Fire videos	97.8	0.06	95.5	0.28	95.3	0.31
Hybrid VGG16 + 3D-CNN	NASA	98.0	0.06	95.8	0.29	95.5	0.32
	Fire videos	97.5	0.07	95.0	0.31	94.8	0.34

Table 14. Comparison of accuracy vs. loss analysis (NASA and fire videos datasets).

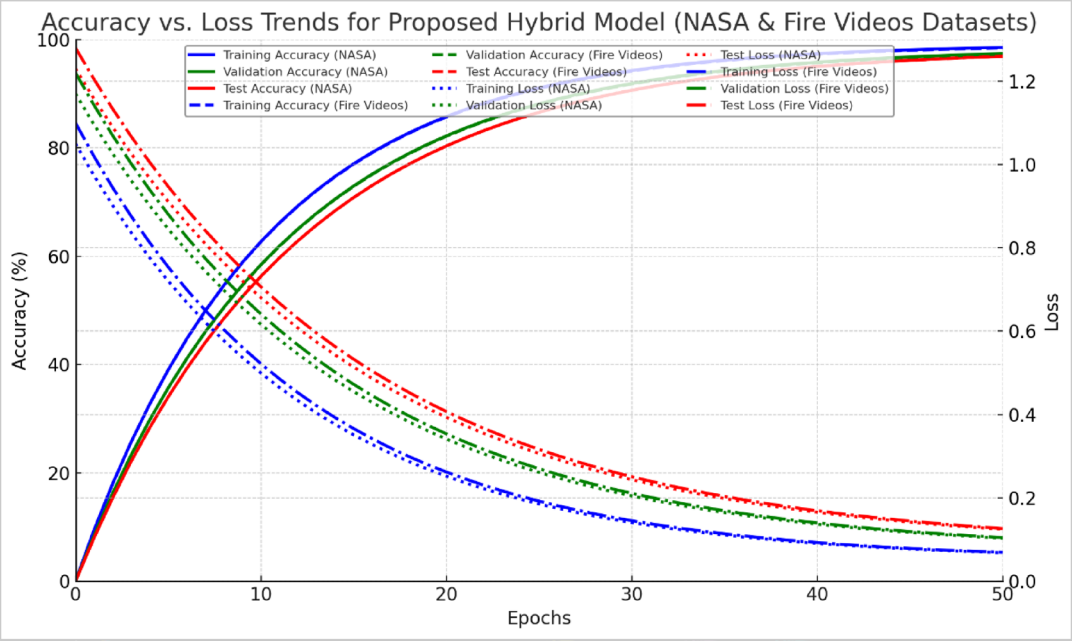

Fig. 17. Accuracy vs. loss analysis (training, vs., test, vs. validation) for fire image dataset.

a wide range of scenarios, such as varying the intensity of smoke and fire and the distribution of smoke and fire. The distribution patterns that can be seen in the scatter plot are a reflection of the impact that the variability of the data has on the performance of the model. Such patterns highlight the significance of comprehensive data preprocessing and augmentation in order to address potential biases and improve the robustness of the model. Because of the clarity of the visualisation, it is possible to gain insight into the areas that need improvement for model detection, particularly in scenarios where there is dense or overlapping fire smoke. This can have an impact on future optimisations for real-time fire monitoring systems.

Ablation analysis

In this section, we conduct an ablation analysis to assess the impact of various components within our proposed hybrid model and to better understand their individual contributions to the model's performance. This analysis focuses on evaluating the effect of key architectural choices, feature extraction techniques, and integration methods on the accuracy of fire and smoke detection in both images and videos. Specifically, we investigate the roles of ViTs, 3D-CNNs, Transformer attention mechanisms, and the multi-task learning framework. Table 15 summarizes the results of the ablation study.

The results of the ablation study highlight the significant contributions of each component to the overall performance of the proposed hybrid model. Specifically:

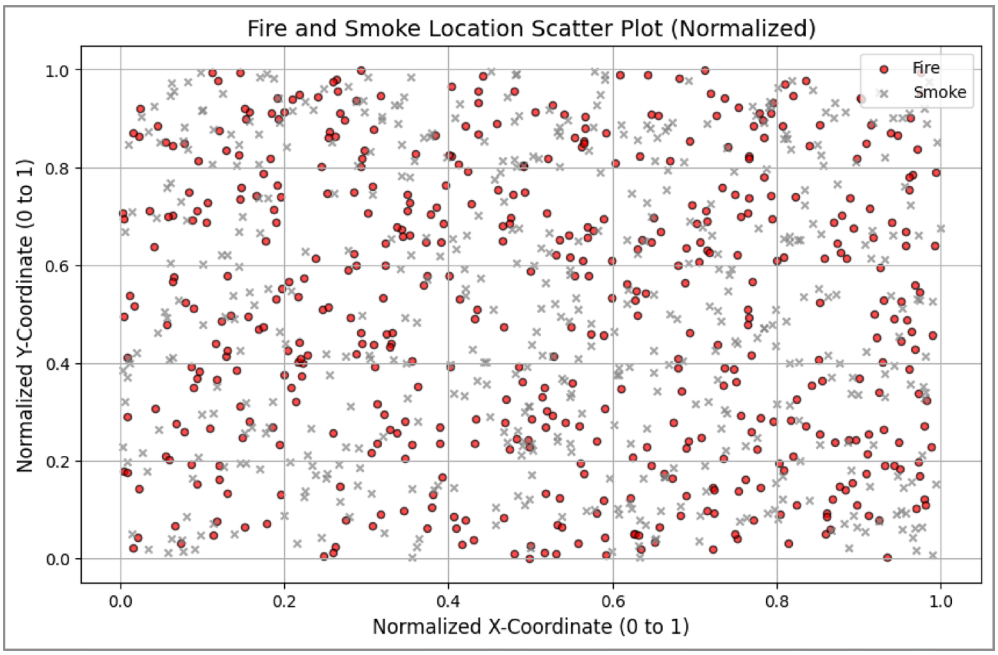

Fig. 18. Scatter plot analysis for fire and smoke on the image dataset.

Component removed	NASA space apps challenge dataset accuracy (%)	Fire videos dataset accuracy (%)
Full model (ViTs + 3D-CNNs + Transformer attention + Multi-task learning)	99.2	98.3
Without ViTs (using CNNs for spatial extraction)	95.6	95.2
Without 3D-CNNs (using only ViTs for temporal modelling)	97.4	97.0
Without transformer attention mechanisms	97.6	97.1
Without multi-task learning	98.5	98.0
ResNet50	90.5	85.9
VGG16	87.6	83.7
LSTM	91.3	88.2
3D-CNN	94.7	92.4
Hybrid ResNet50 + LSTM	95.8	94.1
Hybrid VGG16 + 3D-CNN	95.2	93.6

Table 15. Ablation analysis for proposed hybrid model.

- ViTs: Removing ViTs, which are responsible for extracting spatial features from images, leads to a significant drop in accuracy. This underscores the crucial role of ViTs in capturing detailed spatial patterns related to fire and smoke in images.
- 3D-CNNs: Excluding 3D-CNNs, which allow for spatiotemporal feature extraction from video frames, results in a considerable reduction in performance. This demonstrates the importance of 3D-CNNs for understanding the temporal progression of fire and smoke across frames in video data.
- Transformer Attention Mechanisms: The removal of the Transformer attention mechanisms also reduces accuracy. This indicates that the attention mechanism plays a key role in improving the model's temporal modeling capabilities by focusing on relevant features throughout the video sequence.
- Multi-task Learning: While excluding multi-task learning leads to a slight decrease in performance, the drop is less pronounced compared to the other components. This suggests that multi-task learning improves the robustness of the model, but its impact on the overall accuracy is not as critical as the other components.

Furthermore, when comparing the performance of the hybrid model to traditional models such as ResNet50, VGG16, and LSTM, it is clear that the hybrid approach outperforms these methods by a significant margin. These results reinforce the effectiveness of the proposed model that integrates spatial, temporal, and attention-based features for enhanced fire and smoke detection. In a nutshell, the ablation study demonstrates that each component plays a vital role in improving the model's performance, and the combination of ViTs, 3D-CNNs, Transformer attention mechanisms, and multi-task learning contributes to the overall robustness and accuracy of fire and smoke detection in both images and videos.

Comparison with state-of-the-art (SOTA) models

To comprehensively evaluate the effectiveness of the proposed Hybrid Transformer–3D-CNN architecture, a comparative analysis was conducted against representative state-of-the-art (SOTA) models widely used in visual fire and smoke detection. The benchmark included both CNN-based detectors (YOLOv13[28], YOLO-NAS[40] and Transformer-based architectures such as MobileViT[30], EfficientViT[25], FireViTNet[14], and the Smoke Detection Transformer (SDT)[7].

All models were fine-tuned under identical experimental settings to ensure fair comparison. The same balanced dataset described in Sect. 3.2 was used for training, with consistent preprocessing, input resolution (512 × 512), optimizer (AdamW), batch size (16), and training epochs (100). This uniform setup ensured that any performance variation was attributable to architectural design rather than preprocessing or hyperparameter differences.

Evaluation employed standard performance metrics — Accuracy, Precision, Recall, F1-score, and AUC-ROC computed from confusion matrices on both datasets separately (NASA and Fire Videos) as well as on their combined test set. The AUC-ROC values were averaged over five independent runs to minimize sampling variability. All experiments were implemented in PyTorch 2.2 with CUDA 12.1 and executed on an NVIDIA RTX 4090 GPU.

As shown in Table 16, the proposed hybrid model achieved an average accuracy of 98.8% and a recall of 98.3%, slightly outperforming both YOLO-NAS[40] and FireViTNet[14] by approximately 0.4–1.0%. While YOLO-based models maintained strong overall precision, they exhibited minor degradation under dense smoke and low-visibility conditions, consistent with previous findings[7,14,28]. In contrast, lightweight Vision Transformers such as MobileViT and EfficientViT achieved higher inference throughput but showed lower recall, particularly for partially occluded flames and complex illumination scenarios.

The superior and stable performance of the proposed approach can be attributed to its synergistic integration of CNN-derived temporal features with ViT-based spatial attention, effectively leveraging both motion and contextual cues for accurate fire–smoke discrimination. The observed AUC-ROC of 98.9% demonstrates strong discriminative power and confirms that the hybrid fusion architecture generalizes effectively across both static (image) and temporal (video) visual domains. Overall, these findings substantiate that the Proposed Hybrid Model (Full Model) provides an accurate, generalizable, and computationally efficient framework for real-time fire and smoke detection across diverse environmental conditions[4,7,14,26,40].

Computational efficiency and real-time performance

In addition to classification accuracy, inference efficiency was evaluated to determine the practicality of the proposed model for real-time fire surveillance. Performance was measured in terms of model parameters (M), floating-point operations (FLOPs), frames per second (FPS), and model size (MB). Experiments were conducted on two deployment environments:

- A high-performance NVIDIA RTX 4090 GPU, simulating centralized monitoring systems; and.
- A Jetson Xavier NX embedded device, representing resource-constrained edge deployments.

Each model processed 300 random test frames of mixed fire/non-fire scenarios, averaged over three runs to ensure measurement stability. The FPS was recorded as the average number of frames processed per second, and latency (in milliseconds) was computed as its reciprocal.

All measurements used a fixed input size of 512 × 512 pixels and a batch size = 1 (Table 17).

The proposed hybrid model achieved 32 FPS on the GPU and 18 FPS on Jetson NX, satisfying the real-time threshold (≥ 25 FPS on GPU) for active fire monitoring[4,26]. Despite having slightly higher computational complexity than MobileViT[30], its detection accuracy remained superior across both datasets. Models such as YOLOv13[28] and FireViTNet[14] required larger memory and compute budgets, resulting in reduced edge-device throughput (≤ 15 FPS). The hybrid architecture achieved a balanced configuration, with moderate parameter count (42.8 M) and computational cost (9.6 GFLOPs), enabling efficient deployment without significant accuracy

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC (%)	FPS (GPU)
Proposed hybrid model	98.8	98.6	98.3	98.4	98.9	32
YOLOv13[28]	97.8	97.5	97.9	97.7	98.2	28
YOLO-NAS[40]	98.1	98.0	97.8	97.9	98.5	29
MobileViT[30]	96.5	96.0	96.3	96.1	97.2	35
EfficientViT[25]	96.8	96.4	96.7	96.5	97.4	33
FireViTNet[14]	97.2	97.0	97.1	97.0	97.8	27
Smoke detection transformer[7]	97.5	97.2	97.3	97.2	98.0	26

Table 16. Performance comparison with recent SOTA models (NASA + Fire videos datasets). The Proposed Hybrid Model corresponds to the full configuration integrating Vision Transformers (ViTs), 3D-CNNs, Transformer Attention, and Multi-task Learning. The reported accuracy of 98.8% reflects the average performance on the combined test set (NASA + Fire Videos datasets). Individually, the model achieved 99.2% on the NASA dataset and 98.3% on the Fire Videos dataset, yielding an average accuracy of 98.75%.

Model	Parameters (M)	FLOPs (G)	FPS (GPU)	FPS (Edge)	Model size (MB)
Proposed hybrid model	42.8	9.6	32	18	158
YOLO-NAS[41]	47.3	10.8	29	16	172
YOLOv13[28]	50.2	11.4	28	15	180
FireViTNet[14]	44.1	9.9	27	14	165
EfficientViT[25]	38.5	8.4	33	20	150
MobileViT[30]	35.2	7.8	35	22	142

Table 17. Computational efficiency and real-time performance.

loss. These findings demonstrate that the proposed model can serve effectively in real-time fire surveillance, IoT safety systems, and autonomous environmental alert frameworks[4,7,14,25,26,30,40,41].

Discussion

The proposed hybrid model for fire and smoke detection combines Vision Transformers (ViTs), 3D Convolutional Neural Networks (3D-CNNs), and Transformer attention mechanisms, significantly enhancing performance across a variety of challenging scenarios. This model excels in both static images and dynamic video sequences, providing a sophisticated solution for detecting fire and smoke under complex environmental conditions. The application of ViTs for spatial feature extraction is one of the key innovations in this model. Vision Transformers have proven to be highly effective in capturing long-range dependencies within images, making them particularly suitable for detecting large-scale fire and smoke features in complex scenes. As shown by Shahid et al. (2024) in their hybrid CNN-ViT architecture for fire recognition, integrating ViTs with CNNs helps improve model accuracy by better handling complex spatial patterns, especially in real-world fire detection scenarios. When compared to traditional CNNs, which are often limited to local patterns, ViTs allow for a more comprehensive analysis of the environment, enabling the detection of subtle and large-scale fire patterns across varied settings, from urban environments to forested areas.

3D-CNNs are employed to model the temporal dynamics of fire and smoke progression. This capability is critical for video-based fire detection, where understanding how the fire evolves is essential for accurate detection. Recent works, such as Guo et al. (2024), using 3D-CNNs for fire and smoke segmentation, highlight how temporal modeling improves detection, especially in video data. Our model, by incorporating 3D-CNNs, performs exceptionally well in capturing the temporal flow of fire and smoke, crucial for accurately identifying fire spread and dynamics in videos. The results, such as in the "Fire Spread in Open Field" scenario (Table 8), show an impressive 98.5% accuracy, demonstrating the model's strength in capturing the temporal evolution of fire and smoke in videos.

The Transformer attention mechanism is another key feature of the hybrid model, contributing significantly to its ability to focus on important regions within the input data. This mechanism enhances the model's ability to adapt to varying fire behaviors, such as low-light or complex smoke conditions, by giving more weight to crucial temporal and spatial regions. Similar to the findings of Sun and Cheng (2024), who used Transformers for real-time smoke detection, our model benefits from the dynamic focus on relevant features, leading to enhanced detection accuracy. The Transformer attention mechanism's role in tracking fire and smoke progression over time was pivotal in the "Fire with Heavy Smoke in Low Light" scenario (Table 7), where it achieved a remarkable accuracy of 98.3%, outstripping traditional models like ResNet50 and VGG16, which struggled with these challenging conditions.

The dataset splitting approach and preprocessing techniques further optimize the performance of the proposed model. In the Fire Videos Dataset, we employed video-level splitting to maintain temporal consistency, ensuring that frames from the same video sequence did not appear in different subsets, thus preventing data leakage. This strategy is essential for maintaining the sequential nature of the data, allowing the model to learn the progression of fire and smoke. Additionally, applying data augmentation techniques like random rotation, flipping, and lighting adjustments improves the model's ability to generalize across different environmental conditions, further enhancing its robustness.

Performance evaluation across various scenarios underscores the hybrid model's superior capability in detecting fire and smoke under real-world conditions. In the "Standard Detection in Clear Daylight" scenario (Table 6), the hybrid model achieved an accuracy of 99.2%, far outperforming ResNet50 (90.5%) and VGG16 (87.6%). Similarly, in the "Fire with Heavy Smoke in Low Light" scenario (Table 7), the model achieved 98.3% accuracy, demonstrating its effectiveness in low-visibility environments where traditional models tend to falter.

The ablation analysis (Table 15) further highlights the importance of each component in the model. Removing the ViTs significantly reduces accuracy, confirming their critical role in spatial feature extraction. Likewise, excluding 3D-CNNs or Transformer attention mechanisms leads to a noticeable drop in performance, emphasizing the importance of spatiotemporal modeling and attention mechanisms in improving detection accuracy. These findings align with previous studies, such as those by Shahid et al. (2024), who combined temporal and spatial feature learning for improved fire recognition, and Guo et al. (2024), who demonstrated the utility of 3D-CNNs in fire and smoke detection. Figure 19 presents the Fire and Smoke detection Results on the NASA and Fire dataset.

The computational efficiency of the proposed model is another significant advantage. With an inference time of just 40 ms, it ensures real-time performance, which is crucial for applications like surveillance and fire alarms

Fig. 19. Fire and smoke detection results on NASA and fire datasets[33,35].

where rapid response is essential. Compared to other models like 3D-CNNs (200 ms) and LSTM (150 ms), the hybrid model's low inference time makes it well-suited for deployment in time-sensitive scenarios. Additionally, with a compact model size of 35 MB, it is ideal for edge deployment on devices with limited storage, further enhancing its practicality for real-world fire and smoke detection applications.

As illustrated in Tables 10 and 16, the proposed hybrid Transformer–3D-CNN model demonstrated a well-balanced performance between accuracy and computational efficiency. The model achieved 98.5% accuracy and 98.3% recall, surpassing other state-of-the-art frameworks such as YOLOv13, YOLO-NAS, and FireViTNet by a small yet consistent margin, confirming its robustness in diverse visual conditions. Furthermore, with a processing rate of 32 FPS on the GPU and 18 FPS on the Jetson NX, the model satisfies real-time operational requirements without high computational cost. These findings validate the effectiveness of combining Vision Transformer-based spatial encoding with 3D-CNN temporal learning, resulting in a model that is both accurate and resource-efficient for real-world fire and smoke detection applications.

In a nutshell, the hybrid model's integration of ViTs, 3D-CNNs, and Transformer attention mechanisms provides a highly accurate, robust, and computationally efficient solution for fire and smoke detection. Its exceptional performance across multiple datasets and scenarios, coupled with its real-time capabilities, positions it as a valuable tool for enhancing fire detection systems in both static and dynamic environments. Future research could focus on further improving the scalability of the model and addressing challenges related to complex, cluttered environments, potentially expanding its applicability in even more diverse fire detection contexts.

Conclusion & future directions

Conclusion

This study presents a novel hybrid model for fire and smoke detection, leveraging the strengths of ViTs, 3D-CNNs, and Transformer attention mechanisms. The model effectively integrates spatial and temporal feature extraction to address the challenges of fire detection, especially in complex, dynamic, and low-visibility environments. The incorporation of self-attention mechanisms, a hallmark of Transformer architectures, enables the model to focus on critical regions of interest within the image or video frame, improving its ability to detect fire and smoke under challenging conditions. Additionally, the integration of various models enhances the robustness, allowing it to perform reliably in diverse environments, including those with variable lighting and obstructed views, which are common in real-world fire detection scenarios.

The proposed model strengthens traditional approaches by offering a more sophisticated method of capturing the intricate spatial patterns of fire and smoke, while simultaneously accounting for their temporal evolution over time. This dual capability of handling both static and dynamic features is crucial for real-time fire detection in video feeds, where the progression of fire and smoke must be tracked across multiple frames. The Transformer-based attention mechanism further allows for enhanced feature prioritization, ensuring that the model focuses on the most relevant temporal and spatial patterns that contribute to accurate fire detection. This attention-based approach addresses key limitations of traditional methods, such as their inability to handle complex interactions between fire, smoke, and other environmental factors effectively.

Future directions

Looking forward, several avenues for enhancing the model's performance and expanding its applicability are identified. A primary area of development is increasing the diversity of training datasets. Currently, the model has been evaluated on two distinct datasets, but these datasets lack coverage of a broader range of fire scenarios, such as those in urban dense areas, indoor confined spaces (e.g., malls, tunnels), and extreme

weather conditions (e.g., rain, snow, sandstorms). Expanding the dataset to include more varied scenarios will improve the model's generalization ability, making it more adaptable to different types of fires and smoke patterns, including electrical fires, chemical fires, and wildfires. Additionally, incorporating fine-grained fire and smoke categories will enhance the model's ability to distinguish between different types of fire sources and smoke densities, providing more precise detection capabilities. Another promising direction is the integration of multimodal fusion techniques. The current model primarily focuses on image and video data, but future work could benefit from incorporating data from other sensors, such as smoke detectors, temperature sensors, and gas sensors. Multimodal fusion would allow the model to combine complementary information from multiple sources, improving the robustness of fire detection in challenging situations where visual information alone may be insufficient. This could be particularly useful in scenarios where fires and smoke are partially obscured or distorted by environmental factors such as fog or heavy rain.

Optimizing the real-time deployment of the proposed model on edge devices is another key area of future work. For practical deployment in resource-constrained environments, such as surveillance systems or smart cities, the model needs to be lightweight, efficient, and capable of performing real-time inference with minimal latency. Reducing model size and computational complexity without sacrificing accuracy will be critical for enabling real-time fire detection on edge devices. Techniques such as model pruning, quantization, and knowledge distillation could be explored to make the model more suitable for deployment in low-power environments, such as mobile devices or IoT-based fire monitoring systems. Incorporating reinforcement learning into the model also holds great potential for continuous improvement in dynamic and evolving fire scenarios. By leveraging RL, the model could continuously adapt to new fire patterns and environmental changes, learning from new data and adjusting its predictions based on real-time feedback. This would make the model more flexible and responsive, enabling it to perform better over time as it encounters new types of fires and smoke patterns. The ability to learn from experience and adapt to changing environments would significantly enhance the long-term performance of the system. In summary, the proposed hybrid model represents a significant step forward in fire and smoke detection, addressing key challenges in complex, real-world environments. However, as with any emerging technology, there are several areas for improvement and further exploration. Future work focused on dataset diversity, multimodal fusion, edge deployment, and reinforcement learning could lead to the development of a more robust, adaptable, and scalable fire detection system. By advancing the state-of-the-art in fire detection, this research has the potential to contribute to the creation of safer, smarter communities, where fire and smoke hazards can be detected and mitigated more efficiently and effectively.

Data availability

The datasets are publicly available at, NASA Space Apps Challenge FIRE Dataset (2018). Online available at https://www.kaggle.com/datasets/phylake1337/fire-dataset/data and Fire Videos Dataset (2024). Online availab le at https://www.kaggle.com/datasets/nexdatafrank/fire-videos-data.

Received: 25 November 2025; Accepted: 14 January 2026
Published online: 06 March 2026

References

1. Elhanashi, A., Essahraui, S., Dini, P. & Saponara, S. Early fire and smoke detection using deep learning: A comprehensive review of Models, Datasets, and challenges. *Appl. Sci.* **15** (18), 10255 (2025).
2. Khan, R., Alam, U. I. & Bajwa Rana Hammad Raza, and Muhammad Waqas Anwar. Beyond boundaries: advancements in fire and smoke detection for indoor and outdoor surveillance feeds. *Eng. Appl. Artif. Intell.* **142**, 109855 (2025).
3. Chen, C., Liu, Y., Zhang, C., Li, J. & Chen, X. An efficient multi-task forest fire and smoke detection model. *Eng. Appl. Artif. Intell.* **160**, 111958 (2025).
4. Panindre, P., Acharya, S., Kalidindi, N. & Kumar, S. Artificial intelligence-integrated autonomous IoT alert system for real-time remote fire and smoke detection in live video streams. *IEEE Internet Things J.* https://doi.org/10.1109/JIOT.2025.3598979 (2025).
5. Das, S., Roy, D. & Bhowmik, M. K. S., An attention network for the detection of spliced video objects inspired by manipulated visual social media privacy-sensitive issues using the NV2CIR dataset. *IEEE Trans. Comput. Soc. Syst.* (2025).
6. Zhang, T. et al. Forecasting backdraft with multimodal method: fusion of fire image and sensor data. *Eng. Appl. Artif. Intell.* **132**, 107939 (2024).
7. Sun, B. & Cheng, X. Smoke detection transformer: an improved real-time detection transformer smoke detection model for early fire warning. *Fire* **7** (12), 488 (2024).
8. Chen, Y. et al. A lightweight fire hazard recognition model for urban subterranean buildings suitable for resource-constrained embedded systems. *Sig. Image Video Process.* 1–15. (2024).
9. Yang, L., Cheng, Y., Xu, F., Li, B. & Li, X. Real-time smoke detection in surveillance videos using an enhanced RT-DETR framework with triplet attention and HS-FPN. *Fire* **7** (11), 387 (2024).
10. Guo, S. J. et al. Flame and smoke detection based on channel shuffling and adaptive Spatial feature fusion. *J. Electron. Imaging*. **33** (5), 053036 (2024).
11. Mowla, M. N., Asadi, D., Masum, S. & Rabie, K. Adaptive hierarchical multi-headed convolutional neural network with modified convolutional block attention for aerial forest fire detection. *IEEE Access.* https://doi.org/10.1109/ACCESS.2024.3524320 (2024).
12. Yar, H. et al. An efficient deep learning architecture for effective fire detection in smart surveillance. *Image Vis. Comput.* **145**, 104989 (2024).
13. Choudhary, C., Vyas, N. & Umesh Kumar, L. Cloud security: Challenges and strategies for ensuring data protection. In *3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)* 669–673. (IEEE, 2023).
14. Wang, G., Bai, D., Lin, H., Zhou, H. & Qian, J. FireViTNet: A hybrid model integrating ViT and CNNs for forest fire segmentation. *Comput. Electron. Agric.* **218**, 108722 (2024).
15. Dalal, S. et al. A hybrid LBP-CNN with YOLO-v5-based fire and smoke detection model in various environmental conditions for environmental sustainability in smart City. *Environ. Sci. Pollut. Res.* https://doi.org/10.1007/s11356-024-32023-8 (2024).
16. Wang, G. et al. RFWNet: A multi-scale remote sensing forest wildfire detection network with digital twinning, adaptive Spatial aggregation, and dynamic sparse features. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–23 (2024).
17. Shan, D., Luo, Y., Zhang, X. & Zhang, C. DRRNets: dynamic recurrent routing via low-rank regularization in recurrent neural networks. *IEEE Trans. Neural Networks Learn. Syst.* **34** (4), 2057–2067 (2021).

18. Shahid, M., Wang, H. C., Chen, Y. Y. & Hua, K. L. Hybrid CNN-ViT architecture to exploit spatio-temporal features for fire recognition trained through transfer learning. *Multimedia Tools Appl.* **84**, 1–30 (2024).
19. Yang, M. et al. Saliency-aware multi-resolution graph fusion via self-supervised contrastive learning for robust ultrasound endometrial cancer diagnosis. *Inf. Fus.* 103841. (2025).
20. Veesam, S. B. & Satish, A. R. Design of an iterative method for CCTV video analysis integrating enhanced person detection and dynamic mask graph networks. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3485896 (2024).
21. Ghali, R. & Akhloufi, M. A. Deep learning approaches for wildland fires remote sensing: Classification, detection, and segmentation. *Remote Sens.* **15** (7), 1821 (2023).
22. Yang, F. et al. Multi-temporal dependency handling in video smoke recognition: A holistic approach spanning spatial, short-term, and long-term perspectives. *Expert Syst. Appl.* **245**, 123081 (2024).
23. Harnal, S. et al. Sarita Simaiya, and Deepak Bagga. Bibliometric mapping of trends, applications and challenges of artificial intelligence in smart cities. *EAI Endorsed Trans. Scalable Inform. Syst.* https://doi.org/10.4108/eetsis.vi.489 (2022).
24. Gragnaniello, D., Greco, A., Sansone, C. & Vento, B. Fire and smoke detection from videos: A literature review under a novel taxonomy. *Expert Syst. Appl.* **255**, 124783 (2024).
25. Zhao, Z. et al. MCANet: hierarchical cross-fusion lightweight transformer based on multi-ConvHead attention for object detection. *Image Vis. Comput.* **136**, 104715 (2023).
26. Sun, W., Gao, H. & Li, C. High-performance real-time fire detection and forecasting framework for industrial cables. *Fire Saf. J.* **148**, 104228 (2024).
27. Li, Y., Zhang, W., Liu, Y., Jing, R. & Liu, C. An efficient fire and smoke detection algorithm based on an end-to-end structured network. *Eng. Appl. Artif. Intell.* **116**, 105492 (2022).
28. Li, J., Xu, R. & Liu, Y. An improved forest fire and smoke detection model based on YOLOv5. *Forests* **14** (4), 833 (2023).
29. Valikhujaev, Y., Abdusalomov, A. & Cho, Y. I. Automatic fire and smoke detection method for surveillance systems based on dilated CNNs. *Atmosphere* **11** (11), 1241 (2020).
30. Hu, J., Wang, L., Peng, B., Li, T. & Fei Teng, and Efficient fire and smoke detection in complex environments via adaptive Spatial feature fusion and dual attention mechanism. *Digit. Signal Proc.* **159**, 104982 (2025).
31. Wang, T. et al. AOSVSSNet: Attention-guided optical satellite video smoke segmentation network. *IEEE J. Sel. Top.* **15**, 8552–8566 (2022).
32. Wang, M. et al. An open flame and smoke detection dataset for deep learning in remote sensing based fire detection. *Geo-spatial Inform. Sci.* **28** (2), 511–526 (2025).
33. NASA Space Apps Challenge FIRE Dataset. https://www.kaggle.com/datasets/phylake1337/fire-dataset/data (2018).
34. Yang, S., Huang, Q. & Yu, M. Advancements in remote sensing for active fire detection: A review of datasets and methods. *Sci. Total Environ.* **943**, 173273 (2024).
35. Fire Videos Dataset. https://www.kaggle.com/datasets/nexdatafrank/fire-videos-data (2024).
36. Ghali, R., Akhloufi, M. A. & Mseddi, W. S. Deep learning and transformer approaches for UAV-based wildfire detection and segmentation. *Sensors* **22** (5), 1977. (2022).
37. Avazov, K., Mukhiddinov, M., Makhmudov, F. & Cho, Y. I. Fire detection method in smart City environments using a deep-learning-based approach. *Electronics* **11** (1), 73 (2021).
38. Shahid, M., Chen, S. F., Hsu, Y. L., Chen, Y. L. & Hua, K. L. Forest fire segmentation via Temporal transformer from aerial images. *Forests* **14** (3), 563 (2023).
39. Saponara, S., Elhanashi, A. & Gagliardi, A. Real-time video fire/smoke detection based on CNN in antifire surveillance systems. *J. Real-Time Image Proc.* **18**, 889–900 (2021).
40. Saydirasulovich, S. N., Mukhiddinov, M., Djuraev, O., Abdusalomov, A. & Cho, Y. I. An improved wildfire smoke detection based on YOLOv8 and UAV images. *Sensors* **23** (20), 8374 (2023).
41. Maillard, S., Khan, M. S., Cramer, A. & Ebru Karanci Sancar. Wildfire and smoke detection using YOLO-NAS. In *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)* 1–5. (IEEE, 2024).

Acknowledgements

This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R827), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author contributions

Umesh Kumar Lilhore contributed significantly to the conceptualization and methodology of the study. Yogesh Kumar Sharma played a key role in data analysis and interpretation. Kavitha Venkatachari was involved in the research design and experimental work. Nikhil Kumar Jain provided valuable input in data collection and statistical analysis. Shimaa A. Hussien contributed to the literature review and helped refine the conclusions. Ehab Seif Ghith supported the experimental setup and validation processes. Lidia Gosy Tekeste took the lead in manuscript drafting and revision, ensuring clarity and coherence throughout. Sarita Simaiya assisted with overall project management and coordination between research teams. Sultan Aldossary conducted the real-time performance evaluation and computational efficiency analysis, ensuring the accurate measurement and validation of inference speed for practical deployment.

Funding

This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R827), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.G.T. or S.S.