Wildfire detection using hybrid vision transformer and MobileNetV2 with XAI

Saif Mostasir Rohan[1], Md. Sayem Kabir[1], Tasnim Sultana Sintheia[1], Kazi Tanvir[1], Dipta Gomes[1], Mahfujur Rahman[1] and Akinul Islam Jony[2]

[1]*American International University-Bangladesh, 408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh*
[2]*Open University of Catalonia, Rambla del Poblenou, 154-156, Barcelona, 08018, Spain*

Abstract. Wildfires cause severe ecological and economic damage, and limiting it requires rapid and accurate detection. This study presents an image classification framework for wildfire detection in RGB images from the FLAME 3 dataset, captured by an unmanned aerial vehicle. The preprocessing pipeline applies gamma correction, sharpening, contrast-limited adaptive histogram equalisation (CLAHE) and class-balanced data augmentation. The proposed model combines a vision transformer (ViT) and MobileNetV2 in parallel branches, fusing global contextual features with local spatial features. Explainable artificial intelligence techniques (Grad-CAM, Grad-CAM++ and LIME) are used to visualise the image regions that influence the predictions. The model is compared with nine convolutional and five transformer-based baselines and evaluated with accuracy, Cohen's κ, the Matthews correlation coefficient (MCC), the critical success index (CSI) and related metrics. Trained with the AdamW optimiser, the ViT–MobileNetV2 model achieves a test accuracy of 98.44%, a macro-average F1-score of 0.984 and a recall of 1.000 for the fire class on the FLAME 3 RGB classification subset. Inference takes 10–15 ms per image on an NVIDIA T4 GPU in half precision and 50–70 ms on an x86 CPU with INT8 quantisation. The results apply to the evaluated dataset and experimental protocol.

Keywords: wildfire detection, vision transformer (ViT), MobileNetV2, hybrid deep learning model, explainable AI (XAI), Grad-CAM, LIME, UAV imagery

1. Introduction

In Europe, the heat and drought events of 2018–2022 were accompanied by severe fire seasons: in 2022, almost 270 000 ha burned in Spain, and France recorded its largest burnt area since systematic Copernicus observations began in 2006 [14]. Wildfires are also a major source of carbon emissions; fires on peatlands alone can nearly double global fire-driven carbon emissions [1]. The regional assessment for Europe and Central Asia by the United Nations Office for Disaster Risk Reduction identifies climate change and environmental degradation as principal drivers of disaster risk in the region, including wildfire risk [26].

Conventional detection methods, such as satellite monitoring, watchtower surveillance and ground sensor networks, often lack the spatial detail and timeliness required for rapid fire localisation, especially in dense or remote forests; satellite observations, for example, are limited by cloud cover and revisit intervals [27]. Unmanned aerial vehicles (UAVs) equipped with RGB and thermal cameras can collect high-resolution imagery on demand over terrain that is otherwise difficult to access [15]. Convolutional neural networks (CNNs) are widely used to classify such imagery (table 1), but their decisions are difficult to interpret [23], and ensemble models can have long inference times [6] that complicate deployment on UAVs and edge devices.

0009-0007-1756-8400 (S. M. Rohan); 0009-0001-7022-4187 (Md. S. Kabir); 0009-0007-3309-4348 (T. S. Sintheia); 0009-0001-0612-6101 (K. Tanvir); 0000-0003-3019-6051 (D. Gomes); 0000-0002-3245-8767 (M. Rahman); 0000-0002-2942-6780 (A. I. Jony)
22-46135-1@student.aiub.edu (S. M. Rohan); 22-46985-1@student.aiub.edu (Md. S. Kabir); 22-46039-1@student.aiub.edu (T. S. Sintheia); diptagomes@aiub.edu (D. Gomes); mahfuj@aiub.edu (M. Rahman); ajony@uoc.edu (A. I. Jony)

Received 2025-09-21 Accepted 2026-11-20 Published 2026-11-20

© Copyright for this article by its authors, published by the Academy of Cognitive and Natural Sciences. This is an Open Access article distributed under the terms of the Creative Commons License Attribution 4.0 International (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Vision transformers (ViTs) model long-range dependencies and global context through self-attention and perform well in image recognition, but their memory and computational demands limit their use in real-time and mobile settings [18]. Hybrid architectures that pair a transformer with a lightweight convolutional backbone aim to combine global context with efficient local feature extraction. This study evaluates such a hybrid, a ViT–MobileNetV2 classifier, for wildfire detection in UAV RGB imagery from the FLAME 3 dataset. Explainable artificial intelligence (XAI) techniques, namely Grad-CAM, Grad-CAM++ and LIME, are used to show which image regions drive the predictions. The contributions of the study are as follows:

- Hybrid model: a parallel ViT–MobileNetV2 classifier that fuses transformer-based global features with convolutional local features for wildfire image classification.
- Preprocessing: a pipeline of gamma correction, sharpening, CLAHE and class-balanced augmentation, with data partitioned at the level of the original images so that augmented variants of one image do not appear in more than one subset.
- Benchmarking and ablation: a comparison with nine CNN and five transformer baselines, five optimisers and component-wise ablations on the FLAME 3 RGB classification subset, on which the hybrid model reaches a test accuracy of 98.44%.
- Explainability: Grad-CAM, Grad-CAM++ and LIME explanations of the model's predictions.

Hybrid CNN–transformer models have already been applied to wildfire imagery [5], so the contribution of this study is empirical rather than architectural: a controlled evaluation of a parallel ViT–MobileNetV2 hybrid with concatenation-based feature fusion on the FLAME 3 classification subset.

Section 2 reviews related work, section 3 describes the dataset, preprocessing, model and evaluation protocol, section 4 presents the results, and section 5 concludes the paper.

2. Literature review

2.1. Literature search strategy

Prior work on image-based wildfire detection was identified through a structured search of IEEE Xplore, Scopus, Web of Science, SpringerLink, ScienceDirect, the ACM Digital Library and Google Scholar.

The search used combinations of the following keywords and Boolean operators:

- "wildfire" OR "forest fire" AND ("deep learning" OR CNN OR "vision transformer" OR ViT)
- "wildfire detection" AND ("image classification" OR "smoke detection")
- ("remote sensing" OR UAV OR aerial OR satellite) AND "wildfire" AND "deep learning"
- "wildfire" AND ("explainable AI" OR XAI OR Grad-CAM OR LIME)

The search covered studies published between 2017 and 2025; earlier works were included only if they introduced concepts widely adopted in later research.

Studies were included if they (i) addressed wildfire, forest fire or smoke detection using image data, (ii) employed deep learning models such as CNNs or transformers, and (iii) reported quantitative evaluation metrics. Studies were excluded if they relied solely on non-visual sensor data, lacked experimental validation or did not describe their methods in sufficient detail.

Titles and abstracts were screened first, followed by a full-text review of relevance and methodological quality. The search was not intended as a systematic review; its aim was a balanced selection of CNN-based, transformer-based, lightweight and explainable approaches.

2.2. Deep learning for wildfire detection

Table 1 summarises representative studies.

Table 1: Summary of existing works on wildfire detection and classification.

Article	Model	Dataset	Preprocessing technique	Results (%)	Key contributions / Limitations
Khan et al. [12]	VGG19-based transfer learning	DeepFire (new): 1900 images (950 fire/950 no-fire); train: 1520, test: 380	Online sources; cropped ROIs; resized to 250×250, then 128×128 input	Acc: 95 (prec: 95.7, rec: 94.2)	Contributions: UAV system concept; DeepFire dataset; VGG19 benchmark. Limitation: false positives on visually similar scenes (sunset vs. fire)
Tang et al. [24]	ForestResNet (ResNet50-based)	Custom: 175 internet images; train: 150, test: 25	Random crop (256×256), rotation, horizontal flip, centre crop (224×224), normalisation	Acc: 92	Contribution: ForestResNet for forest condition detection. Limitation: limited diversity of natural forest images
Prakash et al. [19]	RBFN-RAISR; Inception-v3	DeepFire (1900 images); train: 1216, val: 304, test: 380	Augmentation (flip, zoom); resized to 224×224; water wave optimisation for feature selection, noise reduction, enhancement	Acc: 97.55 (F1: 93.33)	Contributions: overview of vision-based methods; RBFN-RAISR model. Limitation: forgetting with transfer learning to new tasks
Ghali, Akhloufi, Jmal et al. [5]	TransUNet (ResNet50-ViT / ViT backbones); MedT (CNN-Transformer backbone)	CorsicanFire: 1135 RGB images & masks; train: 815, val: 111, test: 209	Input 224×224 or 512×512; augmentation: rotation (20°), horizontal flip	F1: 97.7 (Trans-UNet), 96.0 (MedT)	Contribution: first study of transformers for forest fire segmentation. Limitations: TransUNet depends on a large pre-trained backbone; MedT can miss small fires
Ghali, Akhloufi and Mseddi [6]	Ensemble (EfficientNet-B5, DenseNet-201); TransUNet, TransFire, EfficientSeg	FLAME: classification: 48 010 RGB images (30 155 fire); segmentation: 2003 RGB images & masks	Classification: 254×254 input, augmentation (rotation, shear, zoom, shift); segmentation: 512×512 input, augmentation (flip, rotation 15°)	Acc: 85.12 (classifica-tion); F1: 99.9 (seg-mentation)	Contributions: ensemble for classification; transformers for segmentation. Limitation: long ensemble inference time

Table 1 (continued)

Article	Model	Dataset	Preprocessing technique	Results (%)	Key contributions / Limitations
Zhang et al. [28]	FT-ResNet50	FLAME: train: 31 501, val: 7874, test: 8617	Mixup and conventional augmentation; 254×254 input; Adam optimiser, Mish activation, focal loss	Acc: 79.48	Contribution: FT-ResNet50 for UAV fire images. Limitation: lower test than training/validation performance due to domain shift
Guan et al. [8]	DSA-ResNet50 (classification); MaskSU R-CNN (segmentation)	FLAME: classification: 8000 RGB images (254×254); segmentation: fire-labelled images from the classification set	Classification: DSA module; segmentation: Labelme ground truth, U-shaped MaskIoU branch	Acc: 93.65 (classification); mIoU: 82.31, F1: 90.30 (segmentation)	Contributions: DSA module for classification; MaskSU R-CNN for instance segmentation. Limitation: segmentation errors on translucent or occluded flames
Islam et al. [11]	Dual-stream: EfficientNetB7 + ACNet (attention connected network)	FLAME: train: 39 375, test: 3000; DeepFire: train: 1520	Input 224×224; Bayesian optimisation of hyperparameters; Grad-CAM for localisation	Acc: 97.45 (F1: 97.12) on FLAME; 95.97 (F1: 95.54) on DeepFire	Contributions: dual-stream attention model; Bayesian hyperparameter optimisation; Grad-CAM for interpretability. Limitation: scope for further optimisation
Tanvir et al. [25]	Ensemble (ResNet50V2, InceptionV3)	Not reported	Not reported; XAI: LIME, Grad-CAM	Acc: 98	Contributions: ensemble model; XAI for interpretability. Limitation: dataset and preprocessing not reported

Reported classification accuracies range from 79.48% for FT-ResNet50 on FLAME [28] to 98% for a ResNet50V2–InceptionV3 ensemble [25], and transformer-based segmentation models reach F1-scores of 96.0–99.9% on CorsicanFire and FLAME [5, 6]. These figures are not directly comparable, because the studies differ in datasets (FLAME, DeepFire, CorsicanFire and custom collections), data splits and tasks. Common strategies include transfer learning [12, 19, 28], attention modules [8, 11] and ensembles [6, 25]. Reported limitations include confusion between fire and visually similar scenes such as sunsets [12], lower test performance due to domain shift [28], long ensemble inference times [6] and missed small fires [5]. Hybrid CNN–transformer designs have been used mainly for segmentation, for example TransUNet with a ResNet50–ViT encoder and MedT [5], while Islam et al. [11] combined two attention-guided CNN streams for classification. Only two of the reviewed studies applied explainability methods (section 2.3).

2.3. Explainable artificial intelligence in visual recognition

Explainable artificial intelligence aims to make the predictions of complex models understandable to humans. In visual recognition, XAI methods typically identify the image regions that most influence a prediction. Gradient-based methods such as Grad-CAM [23] and Grad-CAM++ [2] produce class-discriminative heatmaps, whereas the model-agnostic LIME [21] fits an interpretable surrogate model around an individual prediction. In wildfire detection, Grad-CAM has been used to localise fire regions [11], and LIME and Grad-CAM to interpret an ensemble classifier [25].

3. Methodology

3.1. Overview

Figure 1 summarises the workflow. The images are preprocessed and augmented and then partitioned into training, validation and test sets. A hybrid model combining a vision transformer and MobileNetV2 extracts global and local features, which are fused for classification. The model is trained with tuned hyperparameters, evaluated with standard performance metrics and interpreted with XAI techniques.

3.2. Dataset description

The FLAME 3 computer vision subset (Sycan Marsh) [9] contains 738 image quartets captured by a UAV during a prescribed burn and curated for fire detection and segmentation. Of these, 622 quartets are labelled fire (actively burning fire) and 116 no-fire (the surrounding forest without fire in the frame). Each quartet comprises a raw RGB image, a raw thermal image, a field-of-view-corrected RGB image and a radiometric thermal TIFF. Only the RGB images were used in this study.

3.3. Data preprocessing

All images are resized to $224 \times 224 \times 3$ pixels to match the input size of the backbone networks. Gamma correction ($\gamma = 1.5$) is then applied to adjust luminance non-linearly and enhance mid-tone detail, after which a convolutional sharpening kernel emphasises edges and CLAHE [20] is applied to the luminance channel in LAB colour space with a clip limit of 2.0 and an 8×8 tile grid to enhance local contrast while limiting noise amplification. To balance the classes and reduce overfitting, each class is expanded to 1500 samples with randomised augmentations: horizontal flips, rotations of up to $\pm 10^{\circ}$, scaling between 80% and 120%, translations of up to 20%, Gaussian noise with a variance of 10–20 and brightness–contrast adjustments within ±20%. Figure 2 shows sample images before and after preprocessing, and the ablation study (section 4.3) quantifies the effect of preprocessing and augmentation.

3.4. Dataset split

Augmentation balanced the two classes, and the dataset was then partitioned at the level of the original images: all augmented variants of a given original image were assigned to the same subset, so that no variants of one image appear in more than one of the training, validation and test sets. The resulting 3000 images (1500 per class) were divided into training (70%), validation (15%) and test (15%) sets, as shown in table 2.

Expanding the 116 original no-fire images to 1500 samples means that many no-fire images share a source scene, so the model may learn scene-specific background features. Partitioning at the original-image level keeps variants of training images out of the test set, but the no-fire test images still derive from a small number of original scenes. Section 4.5 discusses the image regions the model relies on.

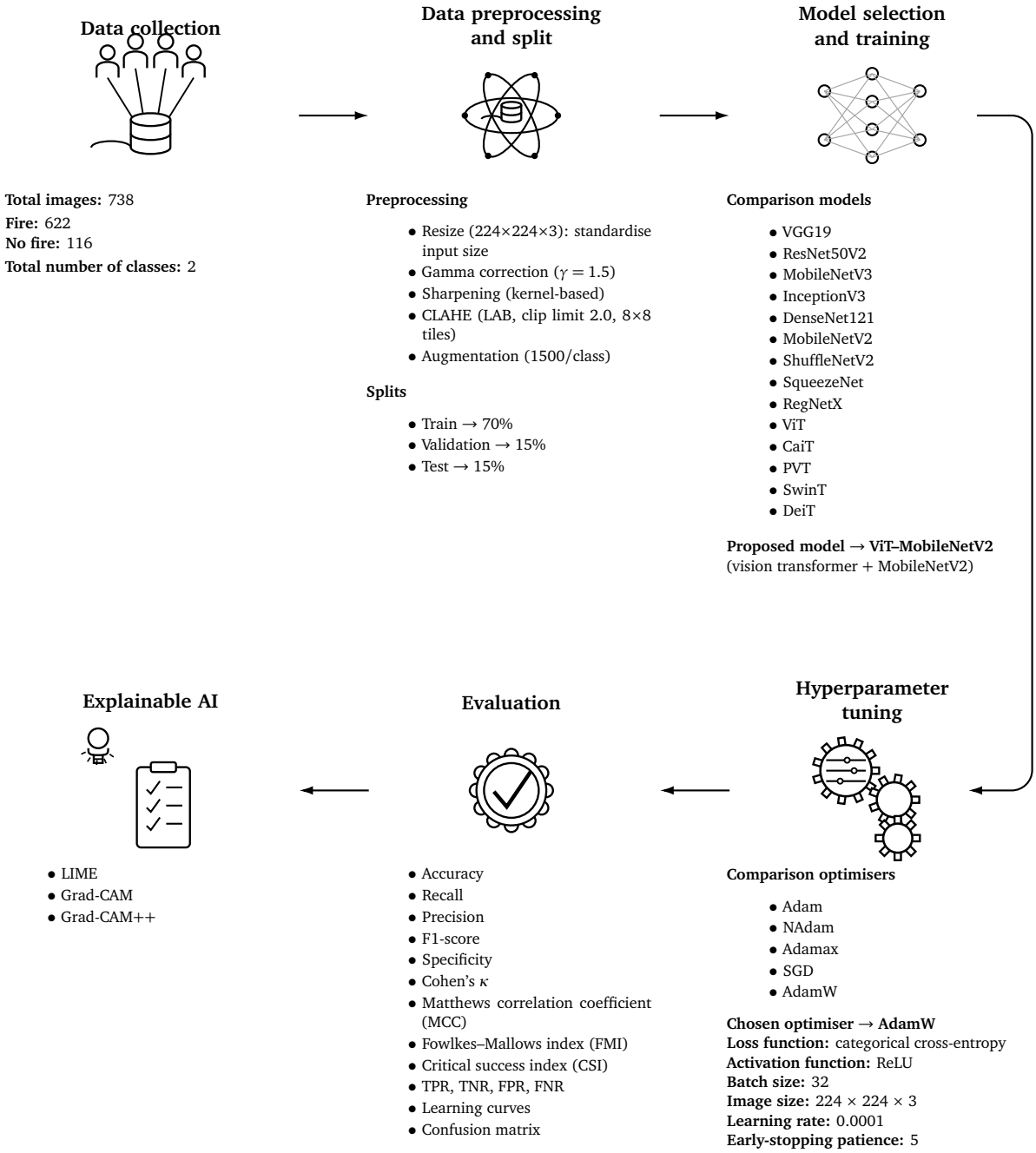


Figure 1: Overview of the methodology.

Table 2
Class-wise dataset split.

Class	Train	Validation	Test	Total
No fire	1050	225	225	**1500**
Fire	1050	225	225	**1500**
Total	**2100**	**450**	**450**	**3000**

3.5. Backbone models

Fire and smoke vary in scale, appearance and spatial extent, so their recognition can draw on both local patterns, such as smoke texture and flame boundaries, and the wider context of terrain and

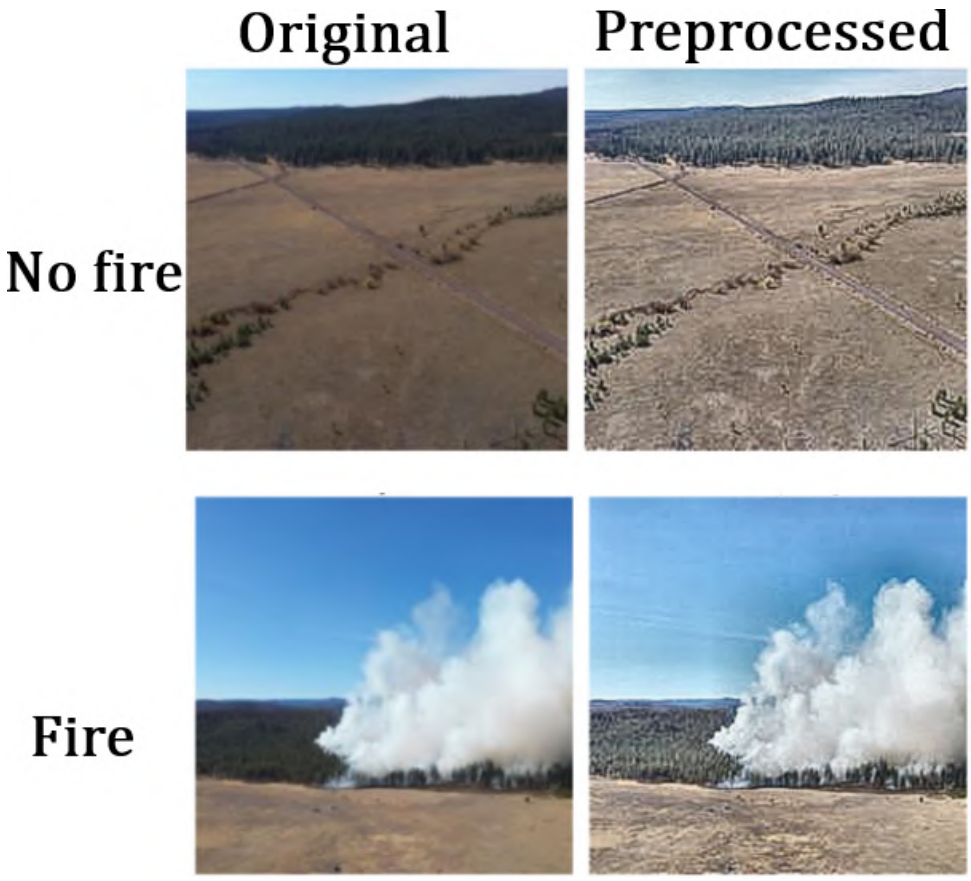


Figure 2: Original and preprocessed sample images for each class.

vegetation. The proposed model therefore combines a transformer backbone for global context with a convolutional backbone for local features.

The vision transformer [3] applies the transformer encoder to images by representing an image as a sequence of fixed-size patches. Each patch is embedded into a feature vector, and multi-head self-attention layers model dependencies across the entire image, which suits wildfire imagery in which smoke plumes and fire regions may be dispersed and vary in scale.

MobileNetV2 [22] is a lightweight CNN designed for efficient feature extraction on devices with limited resources. It uses depthwise separable convolutions and inverted residual blocks with linear bottlenecks to reduce computational cost. Its convolutional layers are suited to extracting local texture and edge features, such as flame boundaries and smoke patterns.

3.6. Proposed hybrid model

The proposed model uses a vision transformer (ViT-Base Patch16/224) and MobileNetV2 as parallel feature-extraction branches (figure 3). The ViT branch outputs the 768-dimensional representation of the `[CLS]` token, and the MobileNetV2 branch outputs a 1280-dimensional feature vector. The two vectors are concatenated into a 2048-dimensional representation and passed to a classification head of fully connected layers with batch normalisation [10], ReLU activations and dropout (rates of 0.5 and 0.3). A softmax layer produces the class probabilities.

3.7. Model configuration and implementation

The model was trained on the free tier of Google Colab with an NVIDIA Tesla T4 GPU (15 GB VRAM) and 12.7 GB of system RAM. The training pipeline was implemented in Python 3 with PyTorch

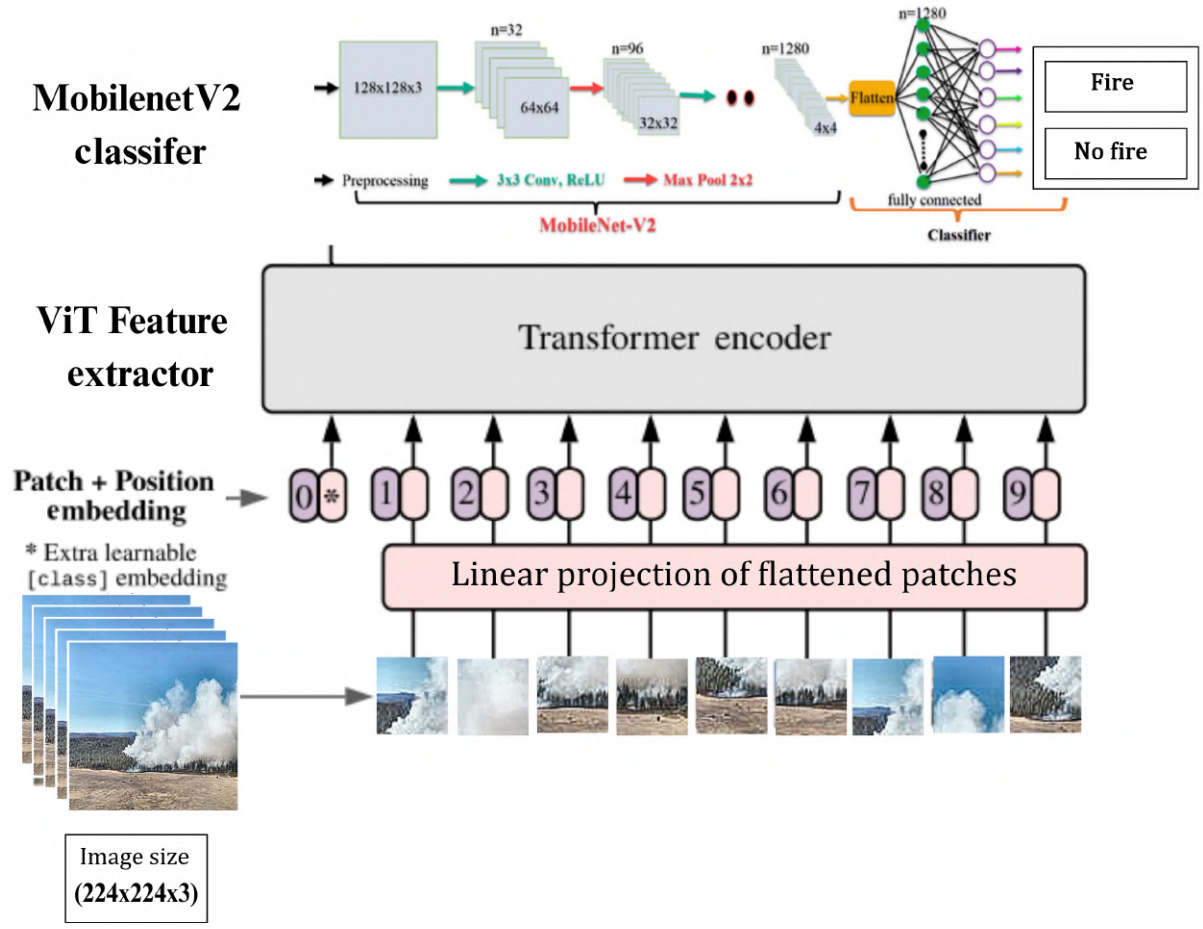


Figure 3: Architecture of the proposed ViT–MobileNetV2 model (the ViT part is adapted from [3]).

and TorchVision, using automatic mixed precision to reduce memory consumption and accelerate computation.

Training ran for up to 100 epochs with early stopping and took approximately 9 s per epoch; it stopped after 39 epochs (figure 5). The AdamW optimiser was combined with a ReduceLROnPlateau learning-rate scheduler, and gradient accumulation was used to simulate larger batches without increasing GPU memory usage.

At inference, the model performs a single forward pass per image. Table 3 reports the latency and throughput on the T4 GPU and on an x86 CPU. Post-training quantisation of the weights from FP32 to INT8 reduced the CPU latency from 120–180 ms to 50–70 ms per image.

Table 3
Inference latency and throughput of the proposed model.

Platform	Latency / throughput	Numerical precision
NVIDIA T4 GPU (FP16)	10–15 ms / 65–100 img/s	Half precision
NVIDIA T4 GPU (FP32)	18–25 ms / 40–55 img/s	Full precision
CPU (x86, INT8)	50–70 ms / 14–20 img/s	8-bit integer post-training quantisation
CPU (x86, FP32)	120–180 ms / 5–8 img/s	Full precision

Figure 4 outlines a conceptual deployment of the model in a UAV-based fire detection and alert system. In this scenario, the trained model would run on an embedded processor mounted on the UAV and classify images from the on-board camera as fire or no-fire. The results would be sent to a central database for monitoring and logging, and a fire prediction would trigger an alert to the

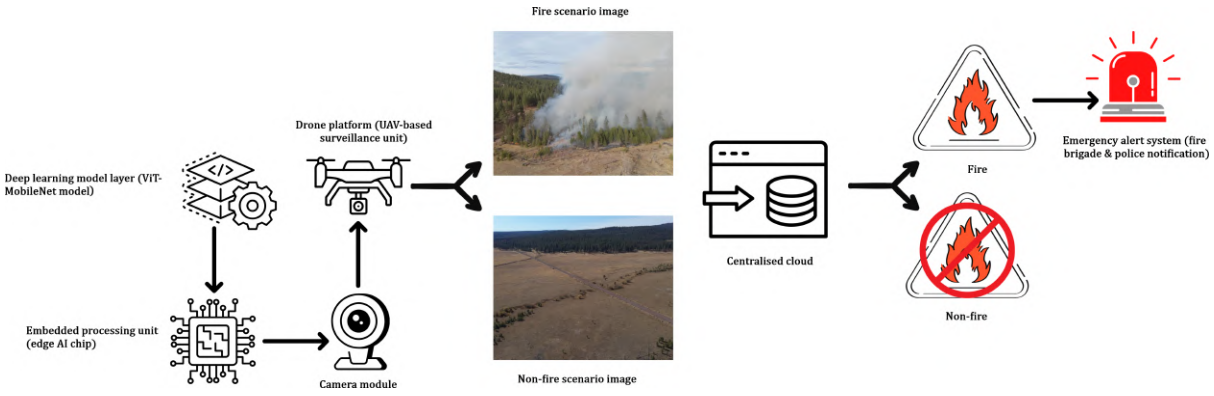


Figure 4: Conceptual UAV-based deployment of the model.

emergency services, such as the fire brigade and police; otherwise, monitoring would continue. This deployment scenario was not implemented or evaluated in this study.

3.8. Training configuration and hyperparameters

The model was trained with a softmax output layer and categorical cross-entropy loss. Five optimisers were compared: Adam [13], NAdam [4], Adamax [13], stochastic gradient descent (SGD) and AdamW [17]. AdamW, which decouples weight decay from the gradient-based update to improve generalisation [17], was used for the final model. Its update rule is

$$\theta_t = \theta_{t-1} - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_{t-1} \right), \tag{1}$$

where η is the learning rate, $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates, λ is the weight-decay coefficient and ϵ is a small constant for numerical stability.

The classification head uses the rectified linear unit (ReLU) activation function [7], which passes positive activations and suppresses negative ones:

$$f(x) = \max(0, x). \tag{2}$$

The categorical cross-entropy loss [16] measures the discrepancy between the predicted class probabilities and the ground-truth labels:

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \tag{3}$$

where C is the number of classes, y_i is 1 if the image belongs to class i and 0 otherwise, and $\hat{y}_i$ is the predicted probability for class i.

Hyperparameters such as the learning rate, batch size, number of epochs and weight decay were tuned empirically on the validation set, and the configuration with the best validation performance was used for the final evaluation. The final configuration used a batch size of 32, a learning rate of 10^{-4} and an early-stopping patience of 5 epochs (figure 1).

3.9. Model evaluation protocol

After the best configuration had been selected on the validation set, the model was evaluated on the held-out test set. The metrics derived from the confusion matrix are accuracy, precision, recall, F1-score and specificity, complemented by two chance-corrected measures, Cohen's κ and the Matthews correlation coefficient (MCC), and by the Fowlkes–Mallows index (FMI) and the critical success index (CSI) for the positive (fire) class. Table 4 lists their definitions.

Table 4
Equations and descriptions of evaluation metrics.

Metric	Equation	Description
Accuracy	$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$	Proportion of correct predictions
Precision	$\text{Precision} = \frac{TP}{TP+FP}$	Proportion of predicted positives that are correct
Recall	$\text{Recall} = \frac{TP}{TP+FN}$	Proportion of actual positives that are detected
F1-score	$\text{F1-score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$	Harmonic mean of precision and recall
Specificity	$\text{Specificity} = \frac{TN}{TN+FP}$	Proportion of actual negatives that are correctly classified
MCC	$\text{MCC} = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	Correlation between predicted and actual labels
Cohen's κ	$\kappa = \frac{p_o - p_e}{1-p_e}$	Agreement beyond chance; p_o and p_e are the observed and chance agreement
FMI	$\text{FMI} = \sqrt{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}$	Geometric mean of precision and recall
CSI	$\text{CSI} = \frac{TP}{TP+FP+FN}$	Overlap between predicted and actual positives

Diagnostic plots complement the scalar metrics. Training and validation accuracy and loss curves show convergence and possible overfitting, and the confusion matrix gives the counts of each error type. F1-score and entropy curves show how the results depend on the decision threshold.

3.10. Explainability methods

Three XAI methods, LIME, Grad-CAM and Grad-CAM++, were used to check whether the predictions of the ViT–MobileNetV2 model rely on image regions that contain fire and smoke rather than on unrelated background.

LIME (local interpretable model-agnostic explanations) [21] explains an individual prediction by fitting an interpretable surrogate model to the behaviour of the trained model in the neighbourhood of the input image. For a given test image, it produces a superpixel-based map of the regions that contributed most to the predicted class. The explanation is obtained as

$$\xi(x) = \arg\min_{g\in G} \sum_i \pi_x(z_i)\big(f(z_i) - g(z_i')\big)^2 + \Omega(g), \tag{4}$$

where f is the trained model, g is an interpretable surrogate model from the class G, z_i are perturbed samples around the input x, z_i' is the binary vector indicating which superpixels are retained in z_i, π_x weights them by their proximity to x, and $\Omega(g)$ penalises the complexity of g.

Grad-CAM [23] was applied to the final convolutional layer of the MobileNetV2 branch. The weight of feature map A^k for class c is the spatially averaged gradient of the class score y^c before the softmax,

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{5}$$

where Z is the number of spatial locations (i, j), and the class activation map is

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right). \tag{6}$$

Grad-CAM++ [2] replaces the uniform average with pixel-wise weights derived from higher-order gradients, which is intended to give better localisation when an image contains several or spatially

extended target regions, such as diffuse smoke. The weights are

$$\alpha_{ij}^{kc} = \frac{\dfrac{\partial^2 y^c}{(\partial A_{ij}^k)^2}}{2\dfrac{\partial^2 y^c}{(\partial A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \dfrac{\partial^3 y^c}{(\partial A_{ij}^k)^3}}, \qquad w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \operatorname{ReLU}\left(\frac{\partial y^c}{\partial A_{ij}^k}\right), \tag{7}$$

and the map is computed as for Grad-CAM, with w_k^c in place of α_k^c.

The resulting maps were inspected qualitatively to assess whether the highlighted regions corresponded to visible fire or smoke. Because Grad-CAM and Grad-CAM++ are computed on the MobileNetV2 branch, they show only the spatial evidence used by that branch; LIME, which perturbs the input image, reflects the full hybrid model.

4. Results and analysis

4.1. Comparison with baseline models

All models were trained and evaluated with the same preprocessing, data split and metrics on the FLAME 3 RGB classification subset. Among the CNN baselines (table 5), MobileNetV2 achieved the highest test accuracy, followed by MobileNetV3 and DenseNet121, and SqueezeNet was the least accurate.

Table 5
Test accuracy of the CNN baselines.

Model	Test accuracy
ResNet50V2	93.21%
VGG19	91.30%
ShuffleNetV2	92.15%
DenseNet121	94.02%
SqueezeNet	90.86%
InceptionV3	93.65%
RegNetX	92.77%
MobileNetV3	94.40%
MobileNetV2	**94.89%**

Among the transformer baselines (table 6), ViT achieved the highest test accuracy, followed by SwinT and DeiT, and PVT was the least accurate. ViT and MobileNetV2, the best models in each group, form the two branches of the proposed hybrid.

Table 6
Test accuracy of the transformer baselines.

Model	Test accuracy
DeiT	93.97%
CaiT	92.56%
SwinT	94.72%
PVT	91.89%
ViT	**95.88%**

With all five optimisers, the ViT–MobileNetV2 model was more accurate than any single baseline (table 7). NAdam and AdamW gave the highest and almost identical test accuracies, and SGD the lowest. AdamW was used for the final model.

Table 7
Test accuracy of the ViT–MobileNetV2 model with different optimisers (the configuration used for the final model is in bold).

Optimiser	Test accuracy
Adam	98.30%
NAdam	98.45%
Adamax	97.90%
SGD	97.50%
AdamW	**98.44%**

4.2. Performance of the proposed model

Tables 8 and 9 report the test-set metrics of the final model, computed from its confusion matrix (figure 7). All 225 fire images were classified correctly (fire recall 1.000), and 7 of the 225 no-fire images were predicted as fire (no-fire recall 0.969, i.e. a false-positive rate of 0.031 for the fire class). The model therefore missed no fires, and all its errors were false alarms. Overall, 443 of the 450 test images were classified correctly (accuracy 98.44%).

Table 8
Class-wise performance of the proposed model.

Class	Precision	Recall (TPR)	F1-score	Specificity (TNR)
No fire	1.000	0.969	0.984	1.000
Fire	0.970	1.000	0.985	0.969

Table 9
Overall performance of the proposed model.

Metric	Value	Metric	Value
Accuracy	0.984	Cohen's κ	0.969
Macro-average precision	0.985	Matthews correlation coefficient	0.969
Macro-average recall	0.984	Critical success index (fire)	0.970
Macro-average F1-score	0.984	Fowlkes–Mallows index (fire)	0.985

4.3. Ablation study

Table 10 shows the effect of removing individual components of the ViT–MobileNetV2 framework. The full model was the most accurate. Using a single branch, omitting feature fusion or omitting the full preprocessing pipeline lowered the accuracy, and omitting class-balanced augmentation lowered it most.

Table 10
Ablation study of the proposed ViT–MobileNetV2 framework.

Variant	Test accuracy
Full model (ViT + MobileNetV2)	**98.44%**
ViT branch only	95.88%
MobileNetV2 branch only	94.89%
Without feature fusion	96.42%
Without full preprocessing	95.73%
Without class-balanced augmentation	94.18%

4.4. Learning curves and diagnostic plots

Figures 5 and 6 show the training and validation accuracy and loss over the 39 training epochs. Training accuracy rose from about 50% to 98% and validation accuracy from about 62% to 98–99%, with both curves levelling off after 15–20 epochs. Training loss fell from about 0.8 to 0.09 and validation loss from about 0.65 to 0.04. The validation accuracy remained at or above, and the validation loss below, the corresponding training values throughout, and the curves do not diverge, so there is no sign of overfitting.

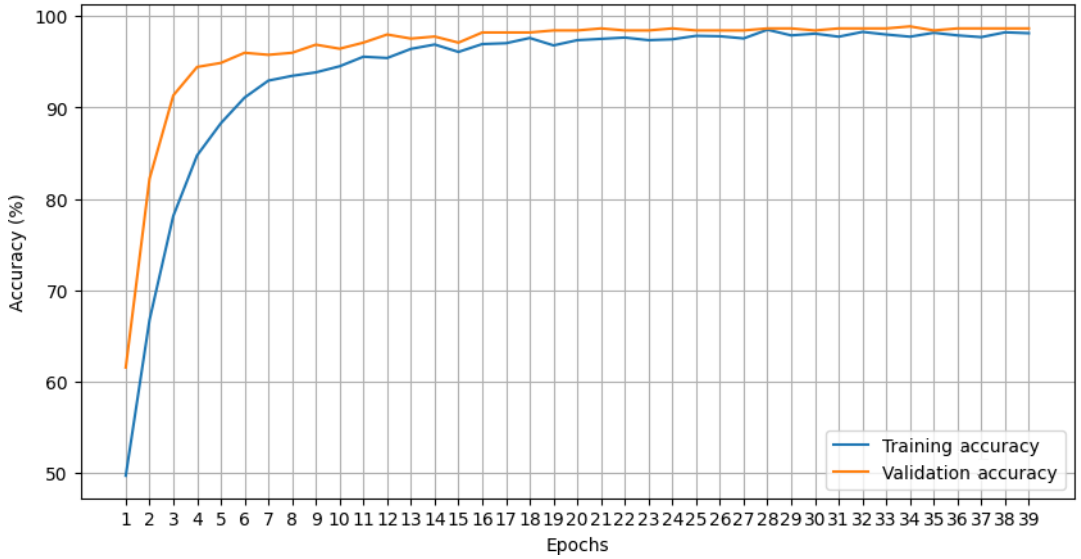


Figure 5: Training and validation accuracy.

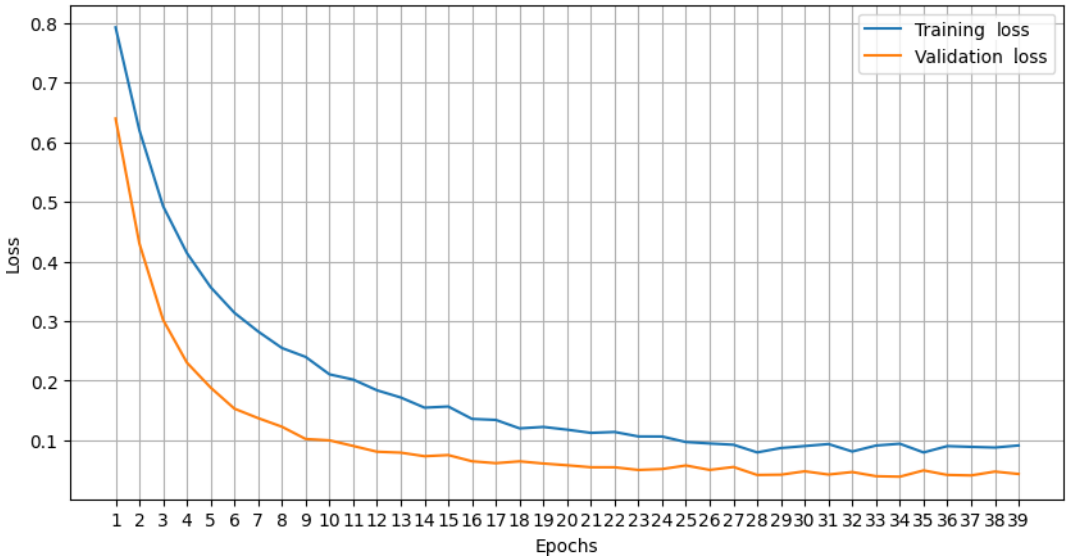


Figure 6: Training and validation loss.

Figure 8 plots entropy against the decision threshold. For both classes, the entropy is close to zero at thresholds of 0 and 1 and reaches a maximum of about 0.69 at thresholds of about 0.5–0.6.

The F1-scores of both classes (figure 9) exceed 0.95 for thresholds between about 0.1 and 0.9 and fall sharply only at the extremes, so the classification performance changes little with the threshold in this range.

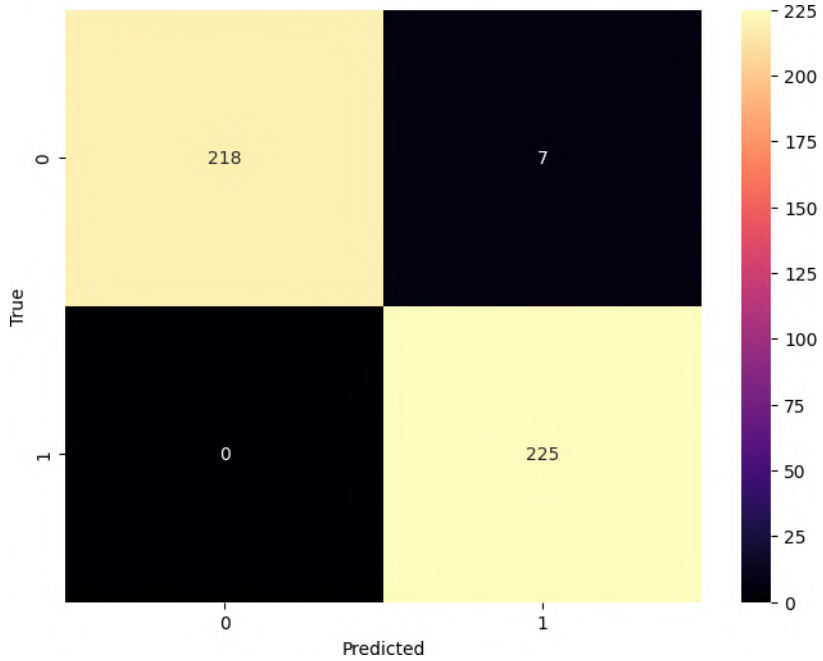


Figure 7: Confusion matrix of the proposed model on the test set (0: no fire, 1: fire).

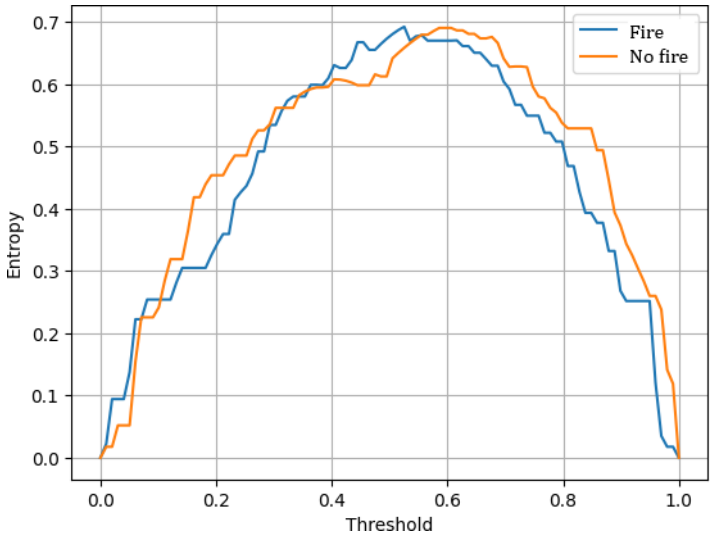


Figure 8: Entropy curves.

4.5. Explainable AI results

Figure 10 compares the LIME, Grad-CAM and Grad-CAM++ explanations for one no-fire and one fire test image. For the no-fire image, LIME highlights large superpixels of bare ground, Grad-CAM produces scattered hotspots around shrubs and surface cracks, and Grad-CAM++ produces a broad activation over the dry ground surface. The no-fire prediction thus appears to rely on the overall ground texture rather than on individual objects; because all no-fire images come from the surroundings of a single burn, it cannot be established whether this cue generalises to other sites. For the fire image, LIME highlights the smoke plume together with the forested slope on the left and patches of ground in the foreground. Grad-CAM and Grad-CAM++ give almost identical maps, with the strongest activation on the forested hillside to the left of the plume and where the smoke crosses the treeline, and weaker activation on the bare ground in the foreground. These are single examples and do not establish the behaviour of the model on the whole test set.

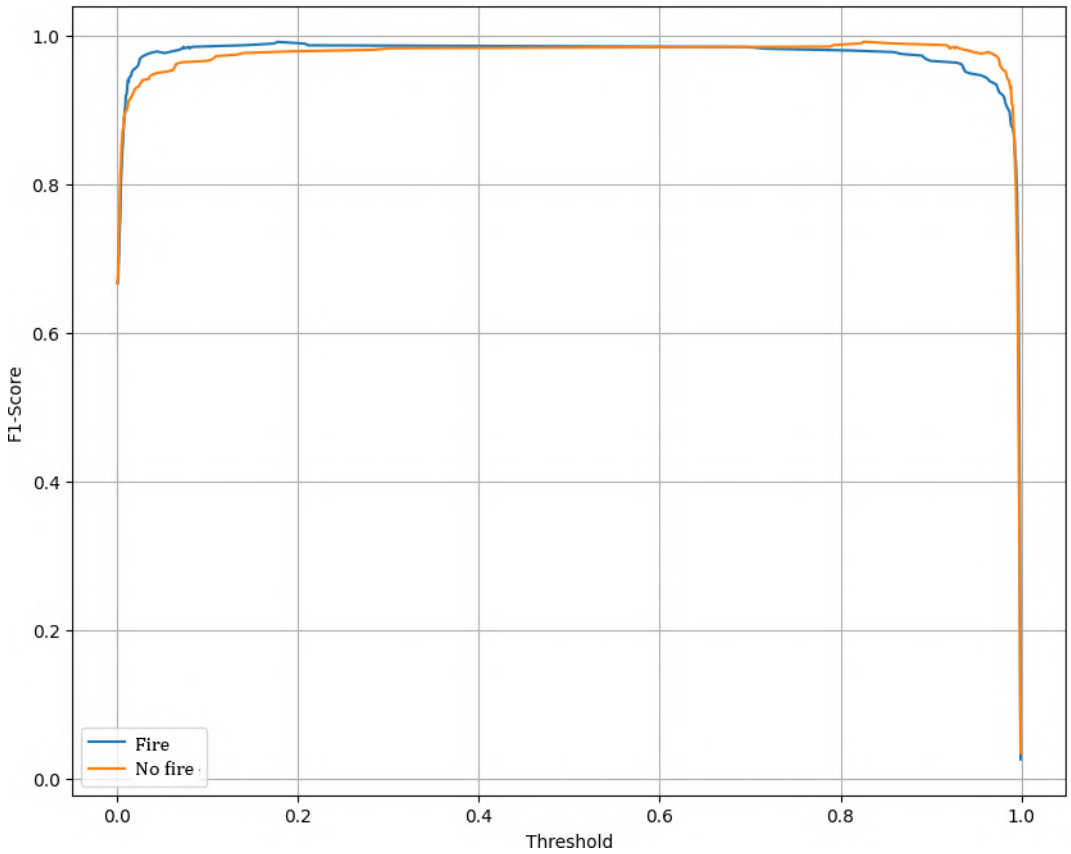


Figure 9: F1-score as a function of the decision threshold.

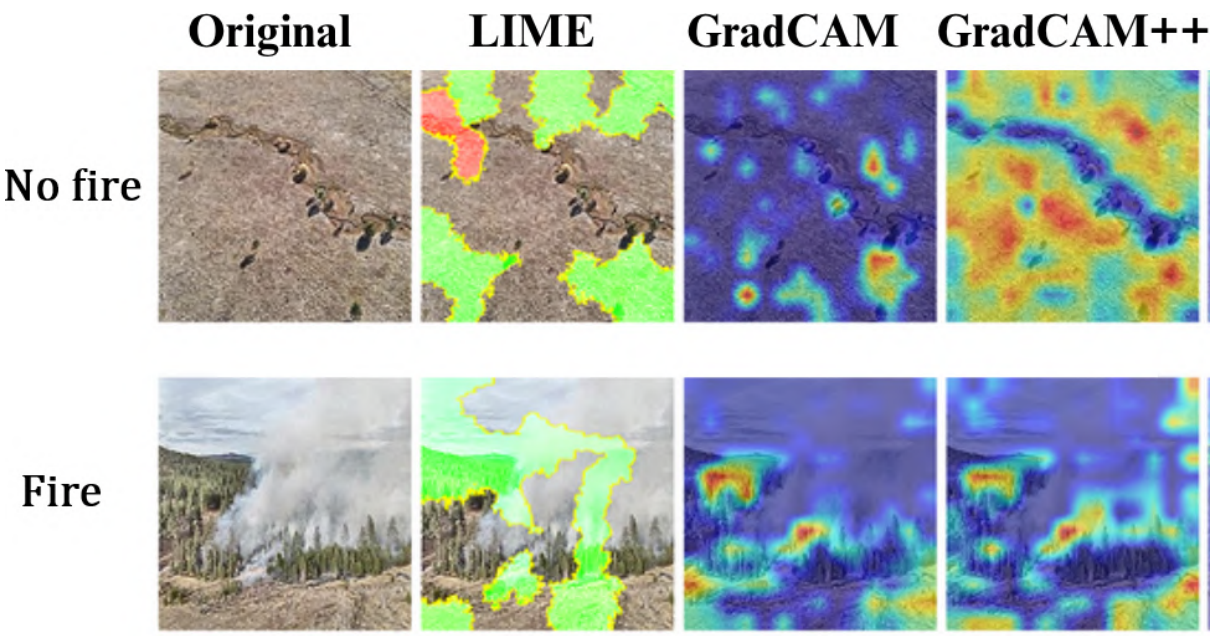


Figure 10: LIME, Grad-CAM and Grad-CAM++ explanations for a no-fire and a fire test image.

5. Conclusion

This study evaluated a hybrid vision transformer and MobileNetV2 classifier for wildfire detection in UAV RGB images from the FLAME 3 dataset. The images were preprocessed with gamma correction, sharpening and CLAHE, and the classes were balanced by augmentation. Trained with AdamW, the

model classified 443 of the 450 test images correctly (accuracy 98.44%, $\kappa = 0.969$, MCC $= 0.969$) and detected all fire images; its seven errors were false alarms on no-fire images. On this test set, the hybrid model was more accurate than nine CNN and five transformer baselines, and removing either branch, feature fusion, the full preprocessing pipeline or class-balanced augmentation lowered its test accuracy. Grad-CAM, Grad-CAM++ and LIME explanations of the example images highlighted smoke, vegetation and ground regions.

The results have several limitations. All experiments used RGB images from a single prescribed burn, the no-fire class was expanded from only 116 original images, and no cross-dataset evaluation or comparison with thermal-based FLAME 3 approaches was performed. Inference was profiled only on a cloud GPU and an x86 CPU, not on embedded UAV hardware. The reported performance therefore applies to the evaluated FLAME 3 RGB subset and experimental protocol.

Future work will extend the evaluation to other datasets, environmental conditions and regions, and will investigate alternative fusion strategies, lightweight transformer variants, knowledge distillation and model compression. Profiling of latency, memory use and power consumption on embedded UAV hardware is also needed, and spatio-temporal modelling of image sequences and the fusion of thermal imagery and environmental data may further improve detection.

References

[1] Baker, S.J., Perry, M.C., Betts, R.A., Schoenecker, J. and Pellegrini, A.F.A., 2025. Spikes in UK wildfire emissions driven by peatland fires in dry years. *Environmental Research Letters*, 20(3), 034028. Available from: https://doi.org/10.1088/1748-9326/adafc6.

[2] Chattopadhay, A., Sarkar, A., Howlader, P. and Balasubramanian, V.N., 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. pp.839–847. Available from: https://doi.org/10.1109/WACV.2018.00097.

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net. Available from: https://openreview.net/forum?id=YicbFdNTTy.

[4] Dozat, T., 2016. Incorporating Nesterov Momentum into Adam. *International Conference on Learning Representations (ICLR) Workshop Track*. Available from: https://openreview.net/pdf/OM0jvwB8jIp57ZJjtNEZ.pdf.

[5] Ghali, R., Akhloufi, M.A., Jmal, M., Souidene Mseddi, W. and Attia, R., 2021. Wildfire Segmentation Using Deep Vision Transformers. *Remote Sensing*, 13(17), 3527. Available from: https://doi.org/10.3390/rs13173527.

[6] Ghali, R., Akhloufi, M.A. and Mseddi, W.S., 2022. Deep Learning and Transformer Approaches for UAV-Based Wildfire Detection and Segmentation. *Sensors*, 22(5), 1977. Available from: https://doi.org/10.3390/s22051977.

[7] Glorot, X., Bordes, A. and Bengio, Y., 2011. Deep Sparse Rectifier Neural Networks. In: G. Gordon, D. Dunson and M. Dudík, eds. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, *Proceedings of machine learning research*, vol. 15. Fort Lauderdale, FL, USA: PMLR, pp.315–323. Available from: https://proceedings.mlr.press/v15/glorot11a.html.

[8] Guan, Z., Miao, X., Mu, Y., Sun, Q., Ye, Q. and Gao, D., 2022. Forest Fire Segmentation from Aerial Imagery Data Using an Improved Instance Segmentation Model. *Remote Sensing*, 14(13), 3159. Available from: https://doi.org/10.3390/rs14133159.

[9] Hopkins, B., O'Neill, L., Marinaccio, M., Afghah, F., Rowell, E., Parsons, R. and Flanary, S., 2024. FLAME 3 Computer Vision Subset (Sycan Marsh). Available from: https://www.kaggle.com/dsv/8480870.

[10] Ioffe, S. and Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: F. Bach and D. Blei, eds. *Proceedings of the 32nd International Conference on Machine Learning*, *Proceedings of machine learning research*, vol. 37. Lille, France: PMLR, pp.448–456. Available from: https://proceedings.mlr.press/v37/ioffe15.html.

[11] Islam, A.M., Masud, F.B., Ahmed, M.R., Jafar, A.I., Ullah, J.R., Islam, S., Shatabda, S. and Islam, A.K.M.M., 2023. An Attention-Guided Deep-Learning-Based Network with Bayesian Optimization for Forest Fire Classification and Localization. *Forests*, 14(10), 2080. Available from: https://doi.org/10.3390/f14102080.

[12] Khan, A., Hassan, B., Khan, S., Ahmed, R. and Abuassba, A., 2022. DeepFire: A Novel Dataset and Deep Transfer Learning Benchmark for Forest Fire Detection. *Mobile Information Systems*, 2022, 5358359. Available from: https://doi.org/10.1155/2022/5358359.

[13] Kingma, D.P. and Ba, J., 2015. Adam: A Method for Stochastic Optimization. In: Y. Bengio and Y. LeCun, eds. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*. Available from: https://doi.org/10.48550/arXiv.1412.6980.

[14] Knutzen, F., Averbeck, P., Barrasso, C., Bouwer, L.M., Gardiner, B., Grünzweig, J.M., Hänel, S., Haustein, K., Johannessen, M.R., Kollet, S., Müller, M.M., Pietikäinen, J.P., Pietras-Couffignal, K., Pinto, J.G., Rechid, D., Rousi, E., Russo, A., Suarez-Gutierrez, L., Veit, S., Wendler, J., Xoplaki, E. and Gliksman, D., 2025. Impacts on and damage to European forests from the 2018–2022 heat and drought events. *Natural Hazards and Earth System Sciences*, 25(1), pp.77–117. Available from: https://doi.org/10.5194/nhess-25-77-2025.

[15] Laghari, A.A., Jumani, A.K., Laghari, R.A. and Nawaz, H., 2023. Unmanned aerial vehicles: A review. *Cognitive Robotics*, 3, pp.8–22. Available from: https://doi.org/10.1016/j.cogr.2022.12.004.

[16] Li, P., He, X., Cheng, X., Qiao, M., Song, D., Chen, M., Zhou, T., Li, J., Guo, X., Hu, S. and Tian, Z., 2022. An improved categorical cross entropy for remote sensing image classification based on noisy labels. *Expert Systems with Applications*, 205, 117296. Available from: https://doi.org/10.1016/j.eswa.2022.117296.

[17] Loshchilov, I. and Hutter, F., 2019. Decoupled Weight Decay Regularization. *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net. Available from: https://openreview.net/forum?id=Bkg6RiCqY7.

[18] Pereira, G.A. and Hussain, M., 2024. A Review of Transformer-Based Models for Computer Vision Tasks: Capturing Global Context and Spatial Relationships. arXiv preprint arXiv:2408.15178. Available from: https://doi.org/10.48550/arXiv.2408.15178.

[19] Prakash, M., Neelakandan, S., Tamilselvi, M., Velmurugan, S., Baghavathi Priya, S. and Ofori Martinson, E., 2023. Deep Learning-Based Wildfire Image Detection and Classification Systems for Controlling Biomass. *International Journal of Intelligent Systems*, 2023(1), 7939516. Available from: https://doi.org/10.1155/2023/7939516.

[20] Reza, A.M., 2004. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement. *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, 38, pp.35–44. Available from: https://doi.org/10.1023/B:VLSI.0000028532.53893.82.

[21] Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp.1135–1144. Available from: https://doi.org/10.1145/2939672.2939778.

[22] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C., 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp.4510–4520. Available from: https://doi.org/10.1109/CVPR.2018.00474.

[23] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE*

International Conference on Computer Vision (ICCV). pp.618–626. Available from: https://doi.org/10.1109/ICCV.2017.74.

[24] Tang, Y., Feng, H., Chen, J. and Chen, Y., 2021. ForestResNet: A Deep Learning Algorithm for Forest Image Classification. *Journal of Physics: Conference Series*, 2024(1), 012053. Available from: https://doi.org/10.1088/1742-6596/2024/1/012053.

[25] Tanvir, K., Kabir, M.S., Anik, Z.H., Hasan, M.R., Tushar, M.T. and Rahman, R.B., 2024. Forest Fire Detection Using Ensemble Deep Learning Model With XAi. *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*. pp.1–6. Available from: https://doi.org/10.1109/COMPAS60761.2024.10796582.

[26] United Nations Office for Disaster Risk Reduction, 2023. *Regional Assessment Report on Disaster Risk Reduction 2023: Europe and Central Asia*. United Nations Office for Disaster Risk Reduction. Available from: https://www.undrr.org/report/rar-2023-europe-and-central-asia.

[27] Yang, S., Huang, Q. and Yu, M., 2024. Advancements in remote sensing for active fire detection: A review of datasets and methods. *Science of the Total Environment*, 943, 173273. Available from: https://doi.org/10.1016/j.scitotenv.2024.173273.

[28] Zhang, L., Wang, M., Fu, Y. and Ding, Y., 2022. A Forest Fire Recognition Method Using UAV Images Based on Transfer Learning. *Forests*, 13(7), 975. Available from: https://doi.org/10.3390/f13070975.