TECHNICAL RESEARCH REPORT

A Comparative Analysis of Deep Learning Architectures for Fire and Smoke Detection in Outdoor Environments

Mohammad Amin Dehmolaee, Peyman Moallem, and Bardia Azami
Department of Electrical Engineering, University of Isfahan, Isfahan, Iran
Research conducted: 2021–2023

Note on this report. This report documents research carried out at the University of Isfahan. It was originally prepared as a journal manuscript and has been revised for presentation as a technical report. The experiments and results are reported as originally obtained; Section 7 discusses the limitations of the study and how it relates to developments in the field since it was conducted.

Abstract

Early detection of fire and smoke is critical for limiting the damage caused by fires. Traffic, surveillance, and satellite cameras provide abundant visual data, which makes vision-based detection a practical first line of defence. However, detecting fire and smoke outdoors is difficult because of the complexity and variability of outdoor scenes. This report presents a comparative study of five widely used image classification networks for outdoor fire and smoke detection: four convolutional neural networks (CNNs), namely ResNet50, ResNet101, RegNetX-1.6GF, and RegNetY-800MF, and one vision transformer, DeiT (Data-efficient Image Transformer). For each network, the ImageNet-pretrained feature extractor is kept frozen and a custom multi-layer perceptron (MLP) classifier is trained on top of it through transfer learning, with separate classifiers for fire and for smoke. To train and evaluate the models, we compiled a dataset of 10,656 images from several sources, including deliberately difficult negative examples, such as objects with fire-like colours and cloud formations that resemble smoke. On a quality-filtered subset of the validation images, the DeiT-based model achieved the best results, with 95.7% accuracy for fire detection and 94.8% for smoke detection. Beyond aggregate metrics, we categorised each model's errors by scene type. This analysis showed that the models have complementary strengths. For example, RegNetX-1.6GF missed fewer aerial smoke images than DeiT, and the ResNets missed fewer small-scale fires than RegNetY-800MF. It also revealed a failure mode shared by all models: they tended to predict fire or smoke whenever firefighters or fire engines appeared in an image, which accounted for up to 35% of false detections. We discuss likely reasons for these behaviours, the limitations of the study, and their implications for building more reliable vision-based fire and smoke detection systems.

Keywords: fire detection; smoke detection; vision transformer; DeiT; convolutional neural networks; transfer learning; error analysis

1. Introduction

Early detection of fire and smoke in outdoor areas is crucial for preventing and limiting the damage caused by wildfires, industrial accidents, and other disasters. Smoke is particularly important because it often appears before flames are visible, so detecting it early allows a prompt response that can stop a fire from escalating. Uncontrolled fires have severe consequences for human life and the environment: they destroy

property and natural resources and pose serious risks to health and safety. According to the National Fire Protection Association (NFPA), approximately 1.3 million fires were reported in the United States in 2019, resulting in an estimated 3,704 civilian deaths and $14.8 billion in property damage [1]. Industrial accidents, such as explosions and fires in chemical plants, can also lead to injuries, fatalities, and environmental contamination. Moreover, the effects of fire extend well beyond the immediate destruction and can include long-lasting environmental and health consequences [2].

Conventional fire detectors, such as heat, flame, and smoke detectors, work well indoors but often fall short outdoors, because they must be close to the fire or smoke to respond, which is impractical over large open areas. Vision-based methods, which can monitor wide areas using existing cameras, have therefore attracted considerable interest. However, the complexity and variability of outdoor scenes make reliable detection difficult. Changing weather and lighting, diverse vegetation, and background elements that resemble fire or smoke can all reduce the accuracy of detection systems.

Deep learning has substantially improved vision-based fire and smoke detection, and two families of architectures are now widely used for image classification: convolutional neural networks (CNNs), which build up their representations from local receptive fields, and vision transformers (ViTs), which use self-attention to model relationships between all parts of an image. Because fire and smoke have no fixed shape and can be spread over large, irregular regions of an image, it is natural to ask whether the global context captured by self-attention gives transformers an advantage over CNNs for this task. This report addresses that question through a comparison of four CNNs (ResNet50, ResNet101, RegNetX-1.6GF, and RegNetY-800MF) and one vision transformer (DeiT), all trained under the same transfer learning setup.

A practical challenge in this work was the lack of a single, representative dataset of outdoor fire and smoke images. We therefore compiled a dataset of 10,656 images from several sources, covering a wide range of scenes, fire and smoke intensities, and environmental conditions, and including difficult negative examples. To improve generalisation, the models were trained on a broad set of images, including indoor scenes and noisy, blurred, or low-quality images. For the final evaluation, however, we used only the validation images that were outdoor scenes and met basic quality requirements, to reflect the intended application (Section 4.2).

Beyond comparing overall metrics, we analysed each model's false positives and false negatives by scene category, to understand where and why the models fail. One notable finding was that all models appeared to associate firefighters and fire engines with fire and smoke, which led to false positives in images that showed firefighters or fire engines but no visible fire or smoke.

The main contributions of this work are as follows:

1. **Comparative evaluation of CNN and transformer backbones.** We compare four CNNs of different depths and design philosophies with a data-efficient vision transformer under an identical transfer learning setup, and report accuracy, precision, recall, F1 score, computational cost, and model size for both fire and smoke detection.
2. **Category-level error analysis.** We categorise each model's false positives and false negatives by scene type (e.g., wildfires, small-scale fires, aerial smoke, thick smoke, and red or orange objects). This reveals complementary strengths among the architectures and a spurious association, shared by all models, between firefighting scenes and fire.

3. **A curated fire and smoke dataset.** We compiled 10,656 labelled images from web searches, frames of YouTube videos, and existing GitHub and Kaggle datasets, including hard negative examples that resemble fire or smoke.
4. **Joint fire and smoke detection.** Separate fire and smoke classifiers share the same frozen feature extractor, so both predictions can be obtained from a single forward pass of the backbone.

The rest of this report is organised as follows. Section 2 reviews related work. Section 3 describes the dataset, the network architectures, and the training approach. Section 4 describes the experimental setup and evaluation metrics. Section 5 presents the results and error analysis, which are discussed in Section 6. Section 7 discusses the limitations of the study, and Section 8 concludes the report.

2. Related Work

2.1 Image-processing approaches

Image processing techniques have been widely used to detect fire and smoke in images and videos. Proposed methods for fire detection include threshold-based [3], colour-based [4, 5], texture-based [6], and Fourier-coefficient-based [7] approaches. Smoke is harder to detect because its visual attributes, such as colour and gradation, are less distinctive and can be difficult to separate from the background. Nevertheless, a range of smoke detection methods has been proposed, including colour-based methods [5], wavelet-based analysis [8], and foreground image accumulation combined with optical flow [9]. Smoke detection in satellite imagery has also received attention, particularly for early wildfire identification [10]. However, these hand-crafted methods often struggle under challenging conditions such as changes in lighting, weather, and background complexity, which has motivated the use of learning-based approaches.

2.2 Convolutional neural networks

CNNs have driven much of the progress in computer vision, from LeNet-5 for digit recognition in the 1990s [11] to AlexNet's breakthrough in large-scale image classification in 2012 [12]. CNNs have since achieved strong results in image classification, object detection, and semantic segmentation. As CNNs became deeper, however, they became harder to train because of the vanishing gradient problem. ResNet [13] addressed this in 2015 with residual (shortcut) connections, which let information flow directly from earlier to later layers and allow much deeper networks to be trained without loss of performance. We use two common ResNet variants. ResNet50 is a 50-layer network that performed exceptionally well on ImageNet when it was introduced [13]; it is built from blocks of convolutional layers with batch normalisation, ReLU activation, and shortcut connections. ResNet101, introduced in the same work, extends this design to 101 layers and is widely used as a deeper, stronger baseline in later work [14].

RegNet, introduced by Radosavovic et al. in 2020 [15], takes a different approach to architecture design. Rather than designing individual networks by hand, the authors progressively simplified a large space of possible network designs and found that good networks follow a simple rule in which the width (number of channels) of each stage increases as a quantised linear function of depth. The result is a family of "regular" networks (RegNets) that can be scaled to different computational budgets while remaining efficient in parameters and computation. RegNetX models use standard residual bottleneck blocks, whereas RegNetY models add squeeze-and-excitation channel attention; the suffix (e.g., 1.6GF or 800MF) indicates the approximate computational cost of the model. In this study we use RegNetX-1.6GF and RegNetY-800MF.

Many studies have demonstrated the potential of CNN-based methods for outdoor fire and smoke detection. Huo et al. proposed 3DVSD, an end-to-end detection network that uses 3D convolutions for real-time video smoke detection with high accuracy [16]. Huang et al. combined the 2D Haar wavelet transform with CNNs in a Wavelet-CNN approach that improved precision and reduced false alarms [17]. Khan et al. proposed SE-EFFNet, a resource-efficient fire detection model that combines a lightweight EfficientNet backbone with stacked autoencoders to reduce computational complexity and mitigate vanishing gradients [18]. Despite their success, CNNs build their representations from local operations and capture global context only indirectly, through many stacked layers. Vision transformers offer an alternative that models global relationships directly.

2.3 Vision transformers

The transformer architecture was introduced by Vaswani et al. in 2017 for natural language processing (NLP) [19]. Its central component is self-attention (specifically, scaled dot-product attention), which lets the model weigh the importance of every element in a sequence relative to every other element and so capture context more effectively than recurrent neural networks (RNNs) or long short-term memory (LSTM) networks. Transformers have since become the foundation of many state-of-the-art NLP models, such as BERT, GPT, and T5, and the standard architecture for tasks such as machine translation, sentiment analysis, and question answering.

Following this success, researchers began applying transformers to computer vision. The first major step was the Vision Transformer (ViT) of Dosovitskiy et al. [20], which divides an image into a sequence of fixed-size patches and processes them with a standard transformer, treating the patches much as words are treated in text. ViTs have since been applied to many vision tasks, including image segmentation [21], medical image analysis [22, 23], and image classification [24–26]. However, as the original authors noted, ViTs require very large training datasets and substantial computation to perform well. Touvron et al. [27] addressed this with the Data-efficient Image Transformer (DeiT), which can be trained effectively on ImageNet alone by combining a carefully designed training recipe (strong data augmentation and regularisation) with knowledge distillation, in which the transformer (the student) learns to reproduce the predictions of a pre-trained teacher network through a dedicated distillation token. This makes transformer-based models more accessible and competitive with CNNs of similar computational cost.

3. Method

3.1 Overview

The goal of this study is to compare how well several widely used image classification architectures detect fire and smoke in outdoor environments. We selected five networks: DeiT, ResNet50, ResNet101, RegNetX-1.6GF, and RegNetY-800MF. All five are well-established, high-performing ImageNet classifiers with publicly available pre-trained weights, and together they cover several contrasts: a vision transformer versus CNNs, a shallower versus a deeper ResNet, and two efficiency-oriented RegNet variants.

For each network, we use the pre-trained feature extractor (the network without its original ImageNet classification layer) and attach a custom three-layer MLP classifier (Figure 1). We train two classifiers with the same architecture, one for fire (flame) and one for smoke, each making a binary decision. Figure 2

illustrates this for the DeiT-based model: a shared feature extractor is followed by separate smoke and fire classifiers.

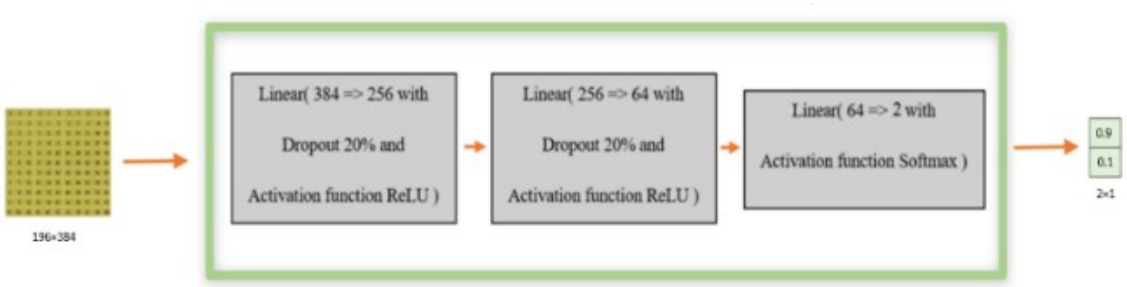


Figure 1. Architecture of the custom classifier attached to each feature extractor (shown for the DeiT backbone).

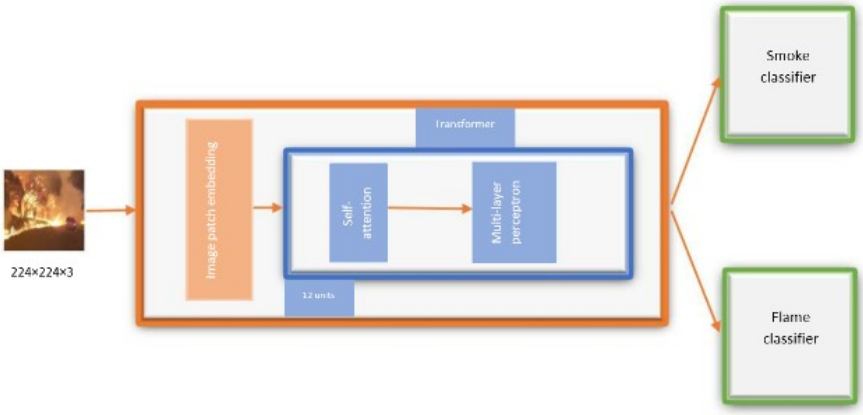


Figure 2. The DeiT-based model for fire and smoke detection: a shared, frozen transformer feature extractor followed by separate smoke and fire classifiers.

3.2 Dataset

Assembling a suitable dataset was one of the main challenges of this work, because at the time we could not find a single available dataset that covered the variety of outdoor fire and smoke scenes we needed. We therefore compiled a dataset of 10,656 images of fire, smoke, and neutral scenes (containing neither fire nor smoke) from several sources: web image searches, frames extracted from YouTube videos, and existing datasets from GitHub and Kaggle. The fire and smoke images cover a wide range of incidents, from large fires and thick smoke to small-scale fires and thin layers of smoke, and the neutral images cover diverse environments.

To make the task realistic and challenging, we deliberately included neutral images containing objects or scenes that resemble fire or smoke, such as objects with fire-like colours and textures and cloud formations that resemble smoke. Figure 4 shows examples from each class. To maximise generalisation, the training data include both indoor and outdoor scenes, as well as images that are blurred, monochrome, or of low quality.

The dataset was split into 8,085 training images (approximately 76%) and 2,571 validation images (approximately 24%), as detailed in Table 1. The validation set is balanced across the three classes, whereas the training set contains roughly twice as many neutral images as fire or smoke images.

Table 1. Number of images per class in the training and validation splits.

Class	Training	Validation	Total
Fire	2,007	857	2,864
Smoke	2,008	857	2,865
Neutral	4,070	857	4,927
Total	**8,085**	**2,571**	**10,656**

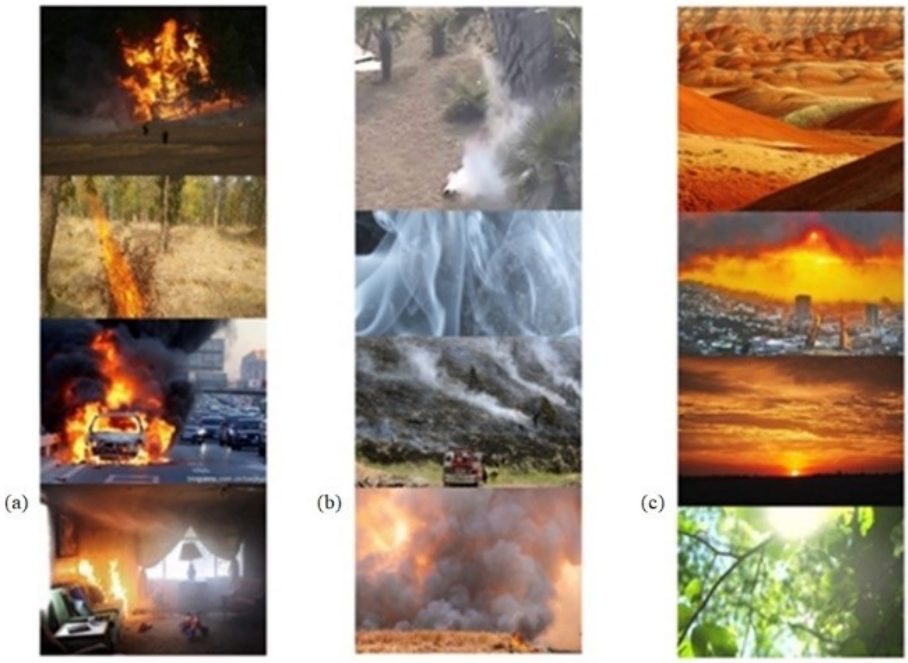


Figure 4. Sample images from the dataset: (a) fire, (b) smoke, and (c) neutral. The neutral examples include scenes with fire-like colours and lighting.

3.3 Why a vision transformer?

CNNs have long been the dominant architecture in computer vision, but they have a structural limitation that is relevant to this task. Each convolutional layer processes only a local neighbourhood of the image, so the receptive field grows only as layers are added, which increases the number of parameters and makes training harder. Techniques such as dilated convolutions can enlarge the receptive field, but the network still builds its understanding of the image from local operations and models relationships between distant regions only indirectly. Vision transformers address this through self-attention. In a ViT, the input image is divided into fixed-size patches, and each patch is embedded as a vector (a token). For the DeiT model used here, a 224×224×3 image is divided into 14×14 = 196 patches of 16×16 pixels. In each self-attention layer, every token is projected into query, key, and value vectors, and the similarity between queries and keys determines how strongly each patch attends to every other patch. As a result, every layer has a global view of the image. This global view is expected to be particularly useful for fire and smoke, which have no fixed shape and can extend irregularly across large parts of an image.

3.4 Transfer learning and classifier

Acquiring a sufficiently large dataset has been a persistent challenge in deep learning. Transfer learning addresses this by reusing a network pre-trained on a large dataset (here, ImageNet) and adapting it to a new task. Computer vision networks typically consist of a feature extractor and a classifier, and transfer learning approaches differ in which of these parts are retrained (Figure 3): (1) retraining the entire pre-trained network on the new dataset; (2) freezing the feature extractor and training only a new classifier; or (3) freezing part of the feature extractor and fine-tuning the remaining layers together with a new classifier. We used the second approach: the weights of each pre-trained feature extractor were frozen, and only the classifier was trained on our dataset. Because all five models share the same trainable component, differences in their performance mainly reflect how well each network's pre-trained features represent fire and smoke, rather than how well each architecture can be fine-tuned.

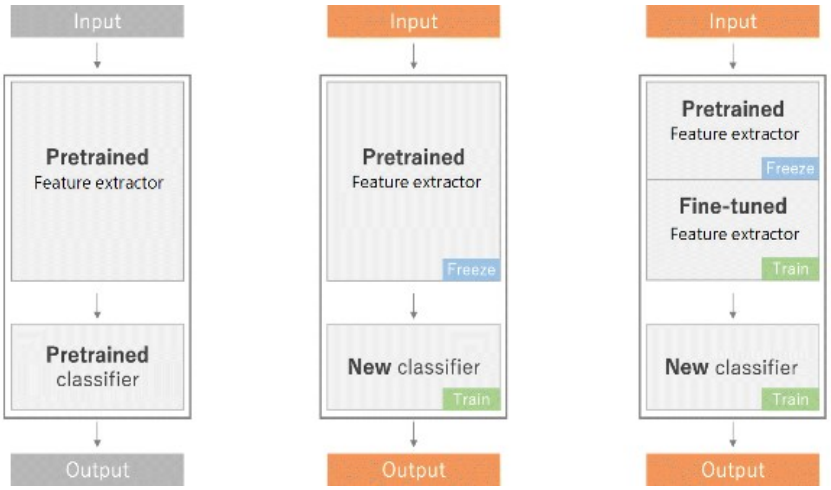


Figure 3. Three common transfer learning strategies. This study uses the second strategy: a frozen feature extractor with a newly trained classifier.

The classifier (Figure 1) consists of three fully connected layers. The first two reduce the feature dimension (384 → 256 → 64 for DeiT) and are each followed by ReLU activation and dropout with a rate of 20%. The final layer maps the 64-dimensional representation to two outputs, and a softmax function converts these into class probabilities that sum to one (fire vs. no fire, or smoke vs. no smoke). For the CNN backbones, the input dimension of the first layer matches the size of the backbone's output features.

3.5 Optimiser

All classifiers were trained with AdamW [29]. Adam [28] adapts the learning rate of each parameter individually and quickly became a standard optimiser for deep networks. However, later studies found that it often generalised worse than stochastic gradient descent (SGD) with momentum, which therefore remained the preferred choice for training many state-of-the-art networks. Loshchilov and Hutter [29] showed that part of this gap arises because L2 regularisation, which is used to prevent overfitting, is not equivalent to weight decay in adaptive optimisers such as Adam. Their variant, AdamW, decouples weight decay from the gradient-based update and applies it directly to the parameters, which improved generalisation across a variety of networks. We chose AdamW for its efficiency and for the more effective regularisation provided by decoupled weight decay.

4. Experimental Setup

4.1 Implementation details

The models were implemented in Python using PyTorch 1.11.0 and trained on Google Colab. Each image was resized to 224×224 pixels and normalised using the ImageNet channel means [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225], which match the preprocessing used to pre-train the backbones. The classifiers were trained with AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1\times10^{-8}$) with a batch size of 64 for 100 epochs.

4.2 Evaluation set

No separate test set was held out. Because the models were never trained on the validation images, the validation set also served for the final evaluation. For each binary task, it contains 1,714 images: 857 fire (or smoke) images as positives and 857 neutral images as negatives. To reflect the intended application, outdoor monitoring with standard colour cameras, the final evaluation used only the outdoor validation images that met a predefined quality threshold; indoor, blurred, monochrome, and low-quality images were excluded. This removed a relatively small number of images, mainly from the fire and smoke classes. Figure 5 shows examples of excluded images. These restrictions apply only to the final evaluation; the training data intentionally retain such images to improve generalisation. The filtered evaluation set nevertheless remains challenging, because it still contains neutral images with fire- or smoke-like appearance and fire and smoke of varying intensity.



Figure 5. Examples of images excluded from the final evaluation, including indoor scenes and low-quality images.

4.3 Evaluation metrics

We evaluate the models with standard binary classification metrics, computed separately for the fire and smoke classifiers. A true positive (TP) is an image correctly classified as containing fire (or smoke); a false positive (FP) is an image incorrectly classified as containing fire (or smoke); a true negative (TN) is an image correctly classified as not containing fire (or smoke); and a false negative (FN) is an image containing fire (or smoke) that the model fails to detect. From these counts we compute accuracy, the overall proportion of correct predictions (Eq. 1); precision, the proportion of positive predictions that are correct (Eq. 2); recall (sensitivity), the proportion of actual positives that are detected (Eq. 3); and the F1 score, the harmonic mean of precision and recall (Eq. 4).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

In a fire detection system, recall reflects how many real incidents are missed, while precision reflects how often the system raises false alarms; both matter in practice. We also report computational cost and model size: the number of floating-point operations (FLOPs) required to process one image, the number of parameters, and the measured inference time.

5. Results

5.1 Training behaviour

To understand how each model learns, we tracked accuracy, precision, recall, and F1 score on both the training and validation sets over the 100 training epochs. Figures 6–9 show these curves for the ResNets and RegNets, each plotted together with the DeiT-based model for reference. For the final evaluation, we selected for each network the checkpoint with the best performance on the validation set.

General trends. Several patterns are common to all models. First, training metrics rise quickly and approach 100% within roughly the first 20 epochs, while validation metrics level off well below them. For the CNNs, validation metrics generally peak within the first 5–10 epochs and then stay flat or decline slowly. The clearest case is smoke detection with RegNetY-800MF, whose validation recall falls from about 95% to about 90% over the course of training (Figure 9). This is a sign of overfitting: even though only the small classifier is trained, it gradually fits the training images at the expense of generalisation. In hindsight, 100 epochs were more than necessary, and early stopping or stronger regularisation would have been appropriate; selecting the best validation checkpoint partly compensates for this. Second, the DeiT-based model shows the same rapid rise in training metrics but no comparable decline in validation performance, and after the first few epochs its validation curves lie above those of the CNNs on almost every metric. Third, the validation curves of the CNNs fluctuate by one to two percentage points from epoch to epoch, so differences of this size between models should not be over-interpreted.

ResNets. Figures 6 and 7 show the training curves of ResNet50 and ResNet101 for fire and smoke detection. The values quoted below are the best values reached on the validation set. For fire detection, ResNet101 outperformed ResNet50 on all metrics except recall, where ResNet50 was more sensitive to true fire events. The largest gap was in precision: 90.4% for ResNet101 versus 88.9% for ResNet50. For smoke detection, the pattern was reversed: ResNet50 outperformed ResNet101 on all metrics except recall (95.6% for ResNet50 versus 96.1% for ResNet101), and the differences between the two networks were larger than in fire detection. On the filtered evaluation set (Section 5.2), however, ResNet50 slightly outperformed ResNet101 in fire detection accuracy and F1 score, so the ranking of these two networks is not stable across data splits.

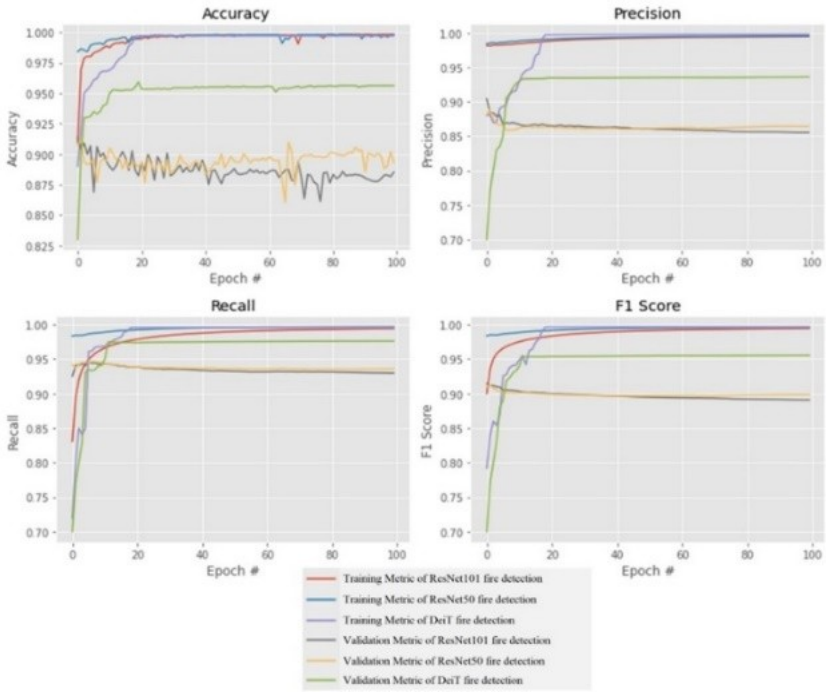


Figure 6. Training and validation metrics of ResNet50 and ResNet101 over 100 epochs for fire detection, with the DeiT-based model shown for reference.

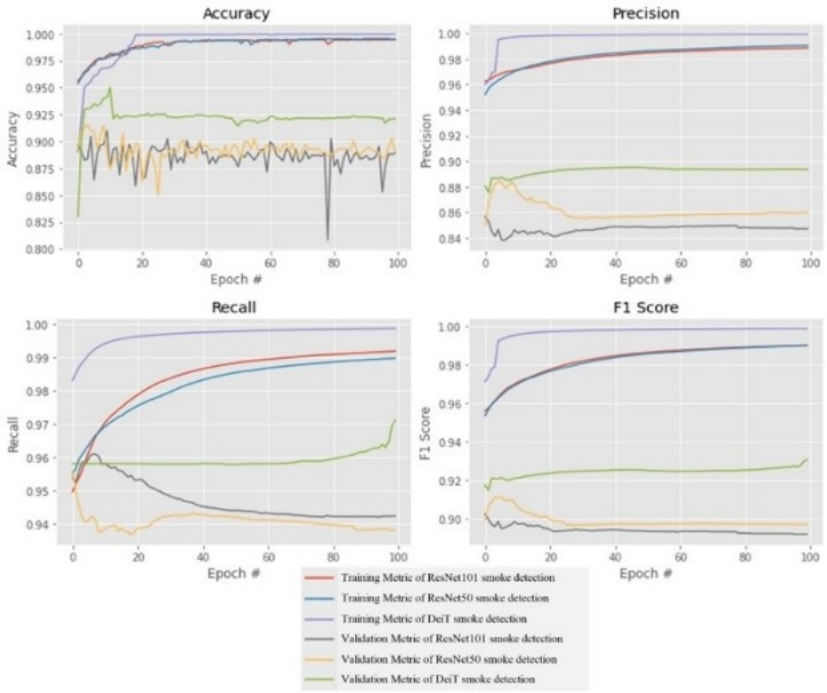


Figure 7. Training and validation metrics of ResNet50 and ResNet101 over 100 epochs for smoke detection, with the DeiT-based model shown for reference.

RegNets. Figures 8 and 9 show the corresponding curves for RegNetX-1.6GF and RegNetY-800MF. For fire detection, RegNetX-1.6GF achieved better validation metrics than RegNetY-800MF for most of training, and its lead was clearest in precision, which reached 90.5%, 1.2 percentage points higher than RegNetY-800MF. Higher precision means fewer false alarms, which matters in applications where false positives are costly. For smoke detection, RegNetX-1.6GF led RegNetY-800MF on all metrics, with a validation accuracy 1.3 percentage points higher, and the gap widened during training as the validation recall of RegNetY-800MF declined (Figure 9).

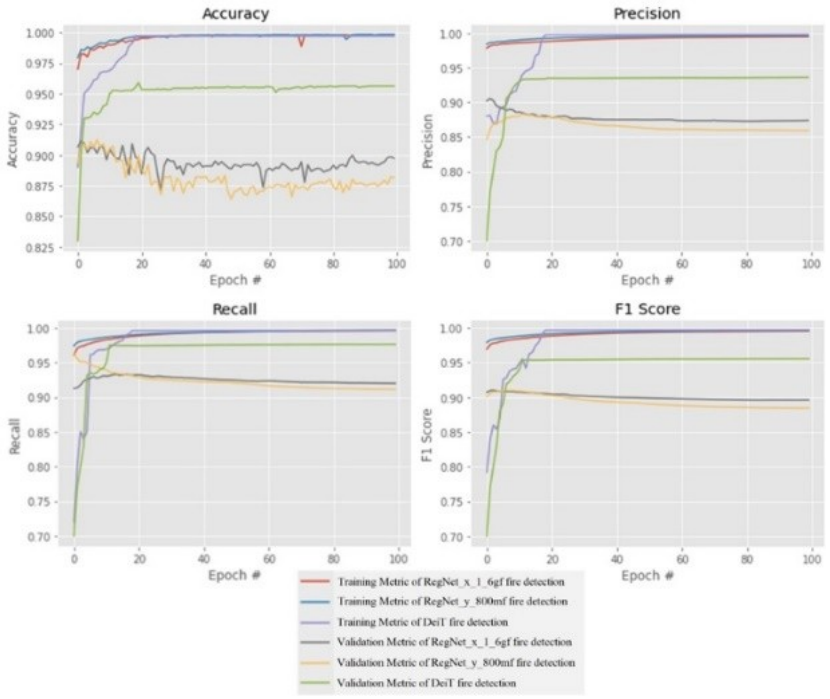


Figure 8. Training and validation metrics of RegNetX-1.6GF and RegNetY-800MF over 100 epochs for fire detection, with the DeiT-based model shown for reference.

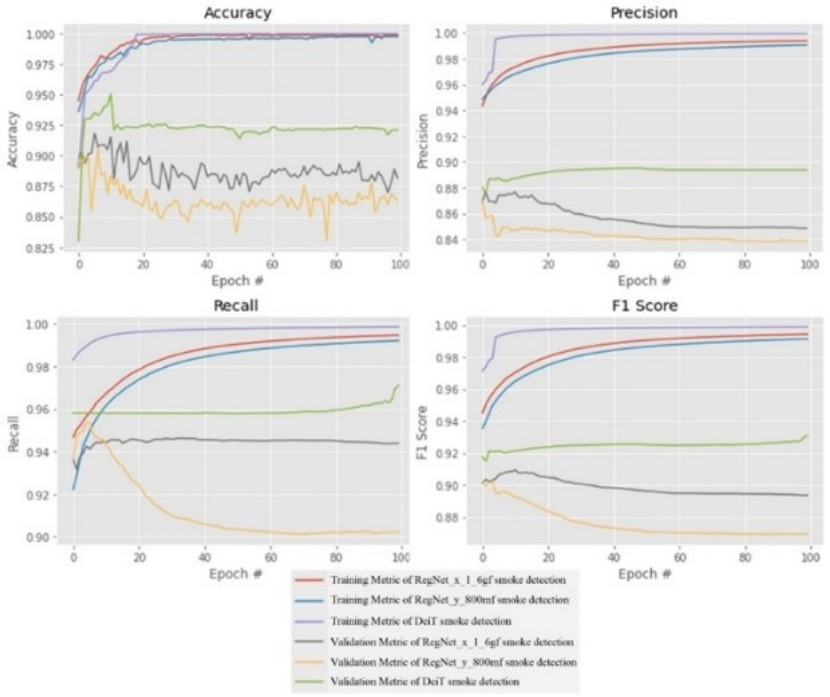


Figure 9. Training and validation metrics of RegNetX-1.6GF and RegNetY-800MF over 100 epochs for smoke detection, with the DeiT-based model shown for reference.

Taken together, no single CNN is best on every metric. Each has advantages on particular metrics, so the most suitable network depends on which kinds of error matter most in a given application. The DeiT-based model, however, maintained higher validation performance than all four CNNs on almost every metric.

5.2 Overall performance

After training, the best-performing checkpoint of each model was evaluated on the filtered evaluation set (Section 4.2). Table 2 summarises the results.

Table 2. Results on the filtered evaluation set. For each model, the first row reports fire detection and the second smoke detection. FLOPs, inference time, and parameter count refer to one model (backbone plus one classifier). Best values in bold.

Model	Task	Accuracy (%)	Recall (%)	Precision (%)	F1 (%)	FLOPs (G)	Time (ms)	Params (M)
DeiT	Fire	**95.7**	**97.9**	**93.4**	**95.6**	8.489	373.4	21.78
	Smoke	**94.8**	**97.7**	**92.1**	**94.8**			
ResNet50	Fire	92.7	95.4	89.9	92.6	8.241	389.8	24.05
	Smoke	91.9	95.4	88.7	91.9			
ResNet101	Fire	92.5	94.5	90.3	92.3	15.7	508.8	43.04
	Smoke	91.4	96.1	89.7	92.8			
RegNetX-1.6GF	Fire	93.0	94.7	91.0	92.8	3.251	372.0	8.53
	Smoke	94.1	97.2	91.1	94.1			
RegNetY-800MF	Fire	92.9	95.0	90.5	92.7	**1.7**	369.7	**5.87**
	Smoke	92.6	94.9	90.3	92.6			

The DeiT-based model achieved the best results on every classification metric for both tasks, with 95.7% accuracy for fire detection and 94.8% for smoke detection. These are 2.7 and 0.7 percentage points higher, respectively, than the second-best model for each task (RegNetX-1.6GF in both cases). The margin for fire detection is substantial, but the margin for smoke detection is small; without repeated runs or confidence intervals, we cannot conclude that DeiT is reliably better than RegNetX-1.6GF for smoke detection. Notably, the deeper ResNet101 did not achieve higher accuracy than ResNet50 on either task, and both RegNet models achieved higher accuracy than both ResNets on both tasks. One anomaly should be noted: for ResNet101 in smoke detection, the reported accuracy (91.4%) is about 1.4 percentage points lower than its precision and recall imply; the values match those recorded during training, and we could not determine the cause.

The models performed better on the filtered evaluation set than on the full validation set, but the improvement was concentrated in the CNNs. As read from the validation curves (Figures 6 and 8), the DeiT-based model reached about 95–96% fire detection accuracy on the full validation set and 95.7% on the filtered set, whereas the CNNs reached at most about 91% on the full validation set but 92.5–93.0% on the filtered set. This suggests that the excluded indoor, low-quality, monochrome, and blurred images (Section 4.2) account for a disproportionate share of the CNNs' errors. In other words, on the harder, unfiltered validation data, DeiT's advantage over the CNNs was larger than Table 2 suggests. The results in Table 2

represent performance under the intended operating conditions. However, because the same validation images were also used to select each model's best checkpoint, they should be regarded as an optimistic estimate of performance on new, unseen data.

5.3 Error analysis

Although the DeiT-based model performed best overall, a closer look at each model's errors gives a more nuanced picture of their strengths and weaknesses. We grouped each model's false positives and false negatives into categories based on image content, which allowed us to identify the conditions under which each network performs well or poorly.

Fire detection. DeiT was particularly reliable on natural and widespread wildfires: only one of its false negatives was a wildfire image. The ResNet-based models, in contrast, struggled with wildfires, especially fires spread over large areas. However, the ResNets performed well on small-scale fires, with only 6 (ResNet50) and 7 (ResNet101) false negatives in this category, compared with 17 for RegNetY-800MF. This may reflect architectural differences between the ResNet and RegNet families; since ResNet50 and ResNet101 performed almost identically despite their difference in depth, depth alone does not appear to explain it. Most false positives involved red or orange objects and lights, or images of the sky. DeiT was the most discerning in these categories, with only 13 false positives, compared with 30 to 46 for the other networks.

Smoke detection. The RegNet models performed well on aerial images of smoke, which are central to wildfire monitoring. RegNetX-1.6GF missed smoke in only 3 aerial images, whereas DeiT missed smoke in 7 aerial images (about 39% of its false negatives). However, the RegNets had difficulty with thick smoke, defined here as smoke dense enough to obscure what lies behind it, whereas DeiT made the fewest errors on both sparse and thick smoke. The ResNets, for their part, missed more smoke in natural settings; whether this makes them better suited to urban scenes would require a direct evaluation on urban smoke images.

A shared failure mode: firefighters and fire engines. A striking pattern was that all networks associated firefighters and fire engines with fire and smoke. This was most evident among the false positives for both tasks: up to 35% of the networks' false detections involved images of firefighters or fire engines with no visible fire or smoke. Figure 10 shows examples of such images that all networks classified as positive.



Figure 10. Examples of images of firefighters and fire engines, without visible fire or smoke, that all networks classified as positive.

6. Discussion

6.1 Why did DeiT perform best?

A plausible explanation for DeiT's overall advantage lies in the nature of fire and smoke. Unlike discrete objects such as vehicles or animals, which have well-defined shapes and parts, fire and smoke are amorphous, transient phenomena, often spread irregularly across large parts of an image. Recognising them reliably may require relating distant regions of the image to one another, which calls for a global view. Self-attention in DeiT computes relationships between all image patches in every layer, whereas CNNs build global context only gradually through many local operations. This explanation is consistent with DeiT's strong results on widespread wildfires and on both thick and sparse smoke.

However, our experiments do not isolate self-attention as the cause of DeiT's advantage, for two reasons. First, the study includes only one transformer, so the comparison is between one specific model and four CNNs rather than between the two architecture families in general. Second, because the feature extractors were frozen, the results largely reflect the quality of each network's ImageNet-pretrained features, which also depend on how each network was trained. DeiT was trained with a modern recipe that includes strong data augmentation and distillation, so part of its advantage may come from its training procedure rather than its architecture. Testing this explanation would require attention-map visualisation and comparisons with more transformers and with CNNs trained using similar recipes.

6.2 Spurious association with firefighting scenes

The association between firefighting scenes and fire is an example of a model relying on a spurious correlation. Firefighters and fire engines frequently appear together with fire and smoke in images of fire incidents, so their presence becomes a convenient but unreliable cue. This matters in practice: a deployed system could raise false alarms whenever emergency vehicles or personnel appear in front of a camera. Possible remedies include adding more negative training images of firefighters and fire engines without fire, and using attention or saliency maps to check which image regions drive the predictions.

6.3 Accuracy versus computational cost

Table 2 also shows that higher computational cost does not guarantee better accuracy. ResNet101 requires almost twice the FLOPs and parameters of ResNet50 but had slightly lower accuracy on both tasks, whereas the two RegNet models achieved higher accuracy than both ResNets with far fewer parameters (8.53M and 5.87M) and FLOPs (3.25G and 1.7G). RegNetY-800MF, the smallest model, reached 92.9% accuracy for fire and 92.6% for smoke with about a quarter of DeiT's parameters and a fifth of its FLOPs. For resource-constrained deployment, such as on edge devices attached to cameras, the RegNets therefore offer an attractive trade-off between accuracy and cost. The measured inference times, however, are almost identical across models (about 370–390 ms, except ResNet101 at 509 ms) despite a roughly fivefold difference in FLOPs. This suggests that the times were dominated by overheads other than the network computation, so they should not be taken as a reliable measure of relative speed.

7. Limitations and Retrospective

7.1 Limitations

The findings of this study should be interpreted in the light of the following limitations:

- **Evaluation protocol.** Results come from a single training run per model, without confidence intervals, so small differences between models (such as the 0.7-point margin in smoke detection) may not be meaningful. No separate test set was held out: the final results come from a quality-filtered subset of the validation set, which was also used to select each model's checkpoint. The reported metrics are therefore likely to be optimistic, both for new data and for unconstrained deployment. The training curves also show that the CNN classifiers began to overfit after the first few epochs (Section 5.1).
- **Dataset.** The dataset combines images from several sources, including frames extracted from videos. Unless all frames from the same video are kept within a single split, near-duplicate images in both the training and evaluation data can inflate measured performance. In addition, the dataset has not been publicly released, which limits reproducibility.
- **Scope of the comparison.** Only one transformer was evaluated, and all backbones were frozen rather than fine-tuned end to end. The conclusions therefore concern frozen ImageNet-pretrained features rather than the full potential of each architecture.
- **Task formulation.** The models perform image-level classification: they indicate whether fire or smoke is present, but not where. Practical systems typically also need localisation, and video-based systems can exploit temporal information, which this study does not use. In addition, each classifier was evaluated against neutral images only, so false alarms caused by the other class (e.g., the fire classifier responding to smoke-only images) were not measured.
- **Error analysis.** The error analysis is qualitative and based on categorising misclassified images. It identifies patterns of failure but does not establish their causes.

7.2 The study in retrospect

This work was carried out between 2021 and 2023, and the field has moved quickly since then. Self-supervised and vision–language foundation models now provide stronger general-purpose image features than the ImageNet-pretrained backbones used here, object detection models are now commonly used to localise fire and smoke in real time, and more public fire and smoke datasets have become available. If we were to repeat this study today, we would (i) fine-tune the backbones in addition to evaluating frozen features; (ii) include several transformer, modern CNN, and foundation-model backbones; (iii) use early stopping and report results over multiple runs with confidence intervals; (iv) hold out a separate test set, split the data at the level of source videos, and evaluate on an external public dataset; and (v) use attention or saliency maps to test the explanations proposed in Section 6. We believe the central observations of this study remain relevant to building reliable fire detection systems: different architectures fail in different and complementary ways, and models can learn spurious cues such as firefighting equipment.

8. Conclusion

This report compared five pre-trained image classification networks, four CNNs and one vision transformer, for detecting fire and smoke in outdoor images, using a frozen-backbone transfer learning setup and a compiled dataset of 10,656 images that includes difficult negative examples. The DeiT-based model achieved the best results on all classification metrics, with 95.7% accuracy for fire and 94.8% for smoke on a quality-filtered subset of the validation set, although its margin over RegNetX-1.6GF for smoke detection was small. A category-level error analysis showed that the networks have complementary strengths: DeiT handled widespread wildfires and both thick and sparse smoke well, the ResNets missed few small-scale fires, and RegNetX-1.6GF detected aerial smoke reliably, while the RegNets as a whole required much less computation. All models also shared a spurious association between firefighting scenes and fire. These findings suggest that the choice of architecture for fire and smoke detection should take into account the deployment scenario and computational budget, not only aggregate accuracy, and that careful error analysis is essential for uncovering failure modes that aggregate metrics hide. Future work should test these findings with fine-tuned backbones, repeated experiments, external datasets, and interpretability methods.

References

[1] Ahrens, M., & Evarts, B. (2020). *Fire loss in the United States during 2019*. National Fire Protection Association, Quincy, MA, USA.

[2] Kim, Y., Knowles, S., Manley, J., & Radoias, V. (2017). Long-run health consequences of air pollution: Evidence from Indonesia's forest fires of 1997. *Economics & Human Biology*, 26, 186–198. https://doi.org/10.1016/j.ehb.2017.03.006

[3] Chen, T.-H., Wu, P.-H., & Chiou, Y.-C. (2004). An early fire-detection method based on image processing. In *Proceedings of the 2004 International Conference on Image Processing (ICIP)* (Vol. 3, pp. 1707–1710). Singapore. https://doi.org/10.1109/ICIP.2004.1421401

[4] Poobalan, K., & Liew, S.-C. (2015). Fire detection algorithm using image processing techniques. In *E-Proceedings of the 3rd International Conference on Artificial Intelligence and Computer Science (AICS 2015)* (pp. 160–168). Penang, Malaysia.

[5] Çelik, T., Özkaramanlı, H., & Demirel, H. (2007). Fire and smoke detection without sensors: Image processing based approach. In *Proceedings of the 15th European Signal Processing Conference* (pp. 1794–1798). Poznań, Poland, 3–7 September 2007.

[6] Dimitropoulos, K., Barmpoutis, P., & Grammalidis, N. (2015). Spatio-temporal flame modeling and dynamic texture analysis for automatic video-based fire detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(2), 339–351. https://doi.org/10.1109/TCSVT.2014.2339592

[7] Liu, C.-B., & Ahuja, N. (2004). Vision based fire detection. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004)* (Vol. 4, pp. 134–137). https://doi.org/10.1109/ICPR.2004.1333722

[8] Chen, T.-H., Yin, Y.-H., Huang, S.-F., & Ye, Y.-T. (2006). The smoke detection for early fire-alarming system base on video processing. In *2006 International Conference on Intelligent Information Hiding and Multimedia* (pp. 427–430). Pasadena, CA, USA. https://doi.org/10.1109/IIH-MSP.2006.265033

[9] Yu, C., Mei, Z., & Zhang, X. (2013). A real-time video fire flame and smoke detection algorithm. *Procedia Engineering*, 62, 891–898. https://doi.org/10.1016/j.proeng.2013.08.140

[10] Sun, Y., Jiang, L., Pan, J., Sheng, S., & Hao, L. (2023). A satellite imagery smoke detection framework based on the Mahalanobis distance for early fire identification and positioning. *International Journal of Applied Earth Observation and Geoinformation*, 118, 103257. https://doi.org/10.1016/j.jag.2023.103257

[11] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. https://doi.org/10.1109/5.726791

[12] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25* (pp. 1097–1105).

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

[14] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1492–1500).

[15] Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollár, P. (2020). Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10428–10436).

[16] Huo, Y., Zhang, Q., Zhang, Y., Zhu, J., & Wang, J. (2022). 3DVSD: An end-to-end 3D convolutional object detection network for video smoke detection. *Fire Safety Journal*, 134, 103690. https://doi.org/10.1016/j.firesaf.2022.103690

[17] Huang, L., Liu, G., Wang, Y., Yuan, H., & Chen, T. (2022). Fire detection in video surveillances using convolutional neural networks and wavelet transform. *Engineering Applications of Artificial Intelligence*, 110, 104737. https://doi.org/10.1016/j.engappai.2022.104737

[18] Khan, Z. A., Hussain, T., Ullah, F. U. M., Gupta, S. K., Lee, M. Y., & Baik, S. W. (2022). Randomly initialized CNN with densely connected stacked autoencoder for efficient fire detection. *Engineering Applications of Artificial Intelligence*, 116, 105403. https://doi.org/10.1016/j.engappai.2022.105403

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30* (pp. 5998–6008).

[20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. arXiv:2010.11929.

[21] Wang, W., Tang, C., Wang, X., & Zheng, B. (2022). A ViT-based multiscale feature fusion approach for remote sensing image segmentation. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5. https://doi.org/10.1109/LGRS.2022.3187135

[22] Raj, R., Mathew, J., Kannath, S. K., & Rajan, J. (2023). StrokeViT with AutoML for brain stroke classification. *Engineering Applications of Artificial Intelligence*, 119, 105772. https://doi.org/10.1016/j.engappai.2022.105772

[23] Manzari, O. N., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B., & Ayatollahi, A. (2023). MedViT: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157, 106791. https://doi.org/10.1016/j.compbiomed.2023.106791

[24] Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., & Ajlan, N. A. (2021). Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3), 516. https://doi.org/10.3390/rs13030516

[25] Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., & Feichtenhofer, C. (2022). MViTv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4804–4814).

[26] Chen, C.-F. R., Fan, Q., & Panda, R. (2021). CrossViT: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 357–366).

[27] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, PMLR 139, 10347–10357. arXiv:2012.12877.

[28] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*. arXiv:1412.6980.

[29] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. arXiv:1711.05101.