Wildfire Detection Using Deep Learning
A Comperative Analysis

Dimitrios Roidis
Christini Tzortzaki

June 2025

1 Introduction

Detecting wildfires quickly and accurately is important to help stop them from spreading and causing damage to the environment, wildlife, and people. Traditional methods like watching from towers or using satellite images can be slow, limited in coverage, or require constant human attention. In contrast, deep learning models can automatically look at images from drones, ground cameras, or satellites and detect signs of fire or smoke in real time.

However, wildfire detection is not easy. Fire and smoke can look very different depending on the lighting, weather, background, and camera quality. For example, a sunset or fog can look like a fire, which can confuse the model. This makes it hard for even advanced computer vision systems to always get it right.

In the past, researchers used simple features like color and edges along with basic machine learning models. These worked okay in simple cases. Later, deep learning models like ResNet and EfficientNet performed much better. More recently, Vision Transformers (ViTs) have shown strong results by understanding the overall image better. Still, many models don't do well when they are trained on small datasets or when they face unfamiliar scenes.

In this project, we test and compare several deep learning models for wildfire detection using a small public dataset. Our main contributions are:

- We compare a **ResNet-18 model trained from scratch** with one that is **pretrained on a larger dataset (ImageNet)** and fine-tuned for wildfire detection.
- We also use a **Vision Transformer (ViT)** model to see how well it can detect fires in complex images.
- To deal with the small dataset, we use **data augmentation**—both common image changes (like flipping and color adjustment) and **synthetic image generation** using a diffusion model to create fake fire images.

Our goal is to find out which models and methods work best for this problem when data is limited and the scenes are complex. By improving the training data and using smart model settings, we aim to build better systems for detecting wildfires in real-world situations.

2 Dataset

Our wildfire dataset [3] is organized into two classes, fire and nofire, with images stored in separate subdirectories under 'train', 'val', and 'test'. In total, we have roughly 400–500 images in each split: about 150–200 "fire" and 250–300 "nofire" per set (e.g. the test set contains 159 fire vs. 251 nofire). All images vary in resolution and scene content—ranging from close-up shots of active flames to wide-angle views of forested

terrain—so the model must learn both local shape cues (flame edges, smoke wisps) and broader context (vegetation, sky). We hold out one portion for validation during training and reserve the test set strictly for final evaluation, ensuring our development choices aren't biased by the data we ultimately report on.

Because 150–200 "fire" examples is relatively modest for training deep models and class imbalance is moderate, training a deep model on the raw data alone risks overfitting and poor generalization to new scenes. To combat this, we will employ targeted data-augmentation techniques during training to synthetically expand the diversity of our samples and help the network learn robust, invariant features across lighting, scale, and background variations, which will be explained thoroughly later.

Figure 1: Non Fire(left) and Fire image(right)

Beyond class imbalance, a number of images in our dataset are genuinely ambiguous even for a human observer. For instance, the intense oranges and reds of a sunset or dawn sky can closely mimic the glow of a distant wildfire Figure 1, leading to false positives, while thick fog or low-lying clouds can resemble the billowing smoke plumes of an active fire, causing false negatives Figure 2. In some cases, thin wisps of smoke against a bright background are nearly imperceptible, and conversely, smoldering embers under heavy vegetation can look like mere shadows. These borderline scenarios highlight why simple color or texture based rules fail and motivate the need for a model that can learn context-aware, invariant features to distinguish true fire signatures from look-alike phenomena.

Figure 2: Non Fire(left) and Fire image(right)

3 Related Work

In recent years, wildfire detection using computer vision and deep learning has emerged as a critical research area due to the growing environmental and economic threats posed by wildfires. Several approaches have been proposed to improve detection accuracy, reduce false positives, and ensure real-time deployability. This section reviews notable work relevant to wildfire image classification, particularly those that have used The Wildfire Dataset or focused on deep learning solutions for class imbalance, model efficiency, and generalization.

El-Madafri et al. [2] introduced The Wildfire Dataset, a diverse collection of labeled "fire" and "no-fire" images sourced from publicly available media and government records. They demonstrated that conventional binary classification models suffer from high false positive rates, especially in the presence of visual confounders like sunsets or clouds. To address this, they proposed a multi-task learning (MTL) framework that combines binary classification with multi-class scene understanding (e.g., distinguishing between fire and similar-looking but non-hazardous scenes). Their MTL model significantly outperformed baseline CNN classifiers such as MobileNetV3, achieving better generalization across challenging, real-world wildfire scenarios.

Building on this dataset, Vuppari et al. [6] investigated the effectiveness of Vision Transformers (ViTs) for wildfire detection. They fine-tuned the ViT-Base-Patch16-224 model using standard data augmentation techniques (resizing, normalization, color jittering) and achieved a test accuracy of 96.1%. This result surpassed those of traditional convolutional models, such as ResNet-50 (88.5%) and VGG-16 (85.3%), highlighting the ViT's strength in capturing global context and modeling long-range dependencies in images—important traits for differentiating smoke or flame textures from complex natural backgrounds.

Figure 3: The process of image tessellation[1].

In parallel, Akagić and Buza [1] proposed LW-FIRE, a lightweight convolutional neural network designed for real-time wildfire image classification. Their work used a different dataset, the Corsican Fire Database (CFDB), but introduced an innovative image tessellation strategy that divides large wildfire images into smaller sub-images to preserve spatial detail and increase training data volume Figure 3. Their best model, LW-FIRE15042, achieved 97.25% testing accuracy while maintaining near real-time inference speeds (25.6 FPS). The compact design of LW-FIRE makes it suitable for deployment on UAVs or embedded systems in wildfire surveillance.

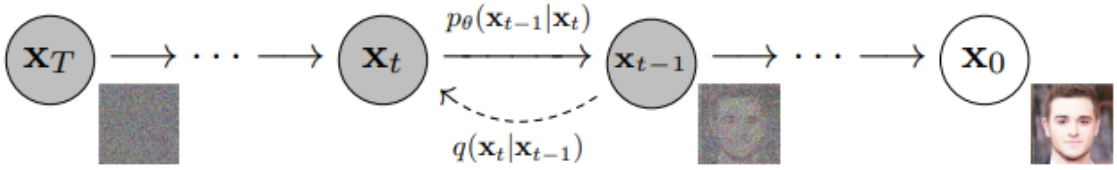

Figure 4: The directed graphical model considered in [4].

Recently, Denoising Diffusion Probabilistic Models (DDPMs) have gained popularity for their ability to generate high-quality, diverse images without the training instability of GANs. First introduced by Ho et al. (2020) [4], diffusion models define a generative process by gradually denoising Gaussian noise through a learned reverse process Figure 4. Their sampling procedure is theoretically grounded in variational inference and Langevin dynamics, allowing for stable and controllable image generation. In this work, we adapt DDPMs for class-conditional data augmentation, using them to generate synthetic wildfire images to address dataset imbalance, a novel application in the domain of wildfire detection.

4 Methods and Approaches

4.1 Data Transformation

In order to tackle the dataset's pathogenic size, we rely on a fully on-the-fly augmentation pipeline to multiply our limited wildfire dataset into a much richer collection of training examples. At the beginning of each epoch, every image is first resized so that its shorter side measures 256 pixels—preserving the original aspect ratio—and then a random 224×224 patch is cropped. This ensures the network learns to identify flames or smoke in varying regions and scales of the scene rather than fixating on a single framing. Next, each crop undergoes a random horizontal flip or a small rotation of up to $\pm 15^{\circ}$, teaching the model to be invariant to camera angle and orientation.

We then perturb the appearance of each image by randomly adjusting brightness, contrast, saturation, and hue within moderate ranges. These photometric transformations simulate different lighting conditions—early-morning sun, heavy smoke haze, or overcast skies—so the classifier cannot rely on a single color palette to detect fire. Finally, the pixel values are converted to tensors and normalized to lie roughly between –1 and 1 in each channel, stabilizing the gradients during training. Because every image in every batch is re-augmented anew on each pass through the data, the model effectively sees thousands of distinct variants over the course of 12 epochs, which dramatically reduces overfitting and builds true robustness to real-world variability.

4.2 Data Generation

In this project, the dataset was significantly imbalanced, with far fewer instances of fire images compared to no-fire images. This imbalance negatively impacted the performance of the classification model, biasing it toward the majority class (nofire). To address this, two approaches were explored:

4.2.1 Image Tiling (Tessellation)

The first method involved tessellating high-resolution images into smaller patches as described in [1] and classifying each patch as either fire or no fire. The rationale was that by breaking down large images into smaller tiles, we could increase the number of training samples, particularly for the minority class, and capture localized features of fire more effectively.

However, this approach proved ineffective in practice. Many of the extracted patches from fire-labeled images did not actually contain visible flames, resulting in significant label noise and inconsistencies. Moreover, in the absence of pixel-level ground truth annotations, we were forced to treat the tessellation process in an unsupervised manner. This further degraded the quality of the dataset, leading to unreliable labels and noisy training signals.

As a result, the classification accuracy for these patches was poor, and the overall dataset quality suffered. Due to these limitations, the tessellation-based augmentation method did not yield meaningful improvements in model performance and was ultimately excluded from the final pipeline.

4.2.2 Denoising Diffusion Probabilistic Model

To handle the class imbalance in the wildfire detection dataset, which contains significantly fewer fire images compared to non fire ones, we employed a Denoising Diffusion Probabilistic Model (DDPM) to synthesize realistic fire images. This approach allowed us to generate new training samples and enhance the classifier's generalization capability without manual data collection.

Following the methodology outlined by Ho et al.(2020) [4], we implemented a UNet-based denoising architecture to reverse a forward diffusion process consisting of 1000 steps of Gaussian noise addition. The model was trained to denoise these corrupted inputs, effectively learning to reconstruct fire images from noise. The model was trained on preprocessed RGB fire images resized and center-cropped to 128×128 pixels, using the Adam optimizer with a learning rate of 1e-4 for 500 epochs.

To quantitatively assess the quality of the generated fire images, we computed the Fréchet Inception Distance (FID) every 10 epochs between the synthetic images and the real fire images from the dataset. Lower FID values indicate higher similarity between the distributions of real and generated images.

Over the course of training, the FID scores progressively improved. The initial FID started very high (2933) due to untrained weights but decreased significantly as training progressed. The best model was observed at epoch 450, where the FID reached 474, indicating a substantial enhancement in visual realism and distributional alignment.

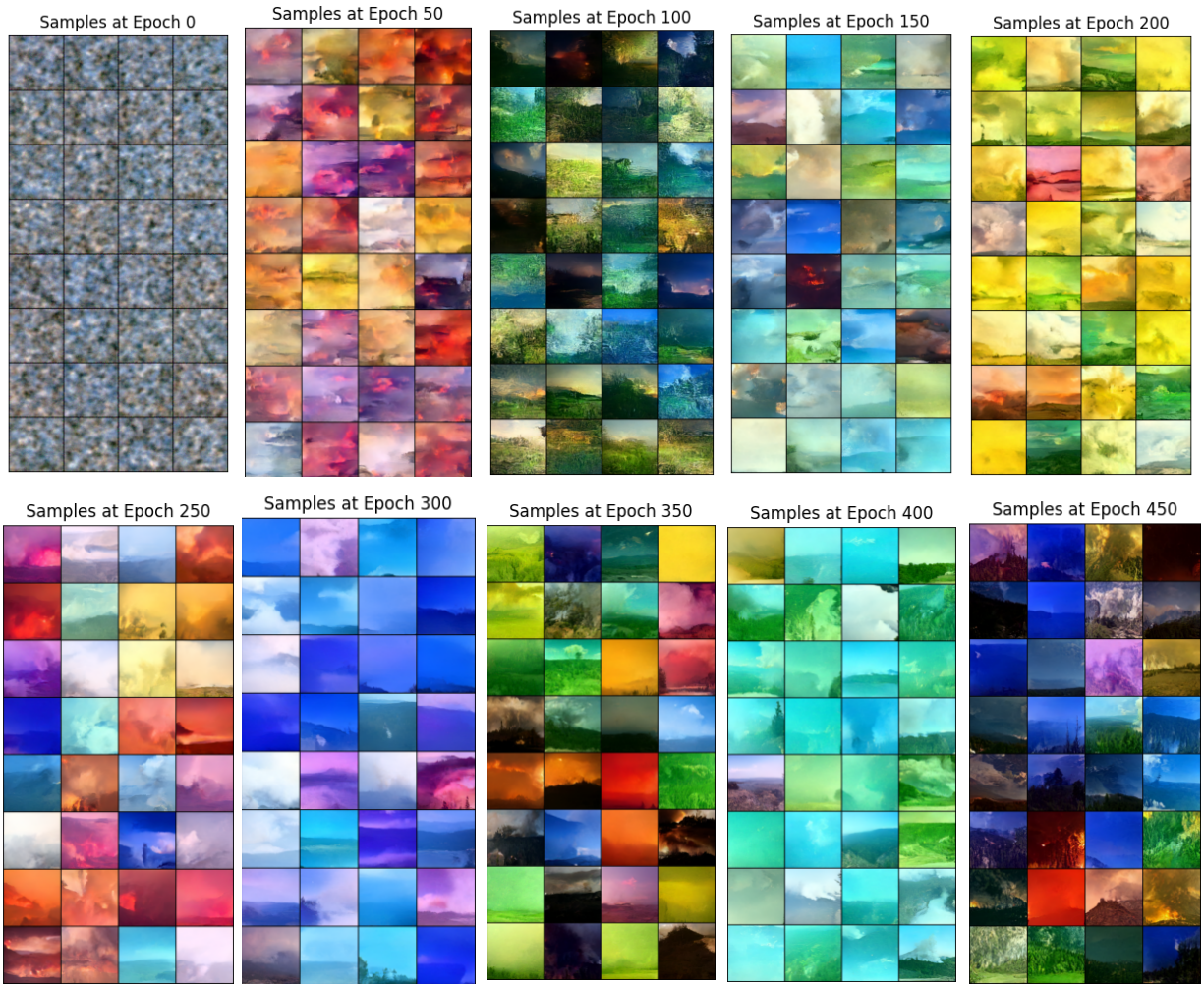

Figure 5: Generated fire images from the diffusion model at every 50 training epochs

The figure 5 illustrates the visual evolution of generated fire images produced by the diffusion model at different training epochs. Each image corresponds to a sample generated at a specific epoch, starting from epoch 0 (untrained model) up to epoch 450 (best-performing model by FID).

In early epochs (e.g., 0–50), the outputs are primarily noise-like and lack coherent structure. As training progresses, the model learns to denoise more effectively, and by later epochs, it generates images that closely resemble realistic fire scenes. This qualitative progression aligns better with the declining Fréchet Inception Distance (FID) scores, confirming that the model converges towards high-fidelity image synthesis.

(a) Original fire images from Wildfire dataset [3]

(b) Generated fire images from diffusion model

Figure 6: Compare fire images

4.3 Models

We evaluate three core architectures for wildfire detection: a ResNet-18 trained from scratch, a ResNet-18 pre-trained on ImageNet and fine-tuned end-to-end, and a Vision Transformer (ViT) [5].

For the "from-scratch" model, we instantiate ResNet-18 with randomly initialized weights and replace its 1000-way ImageNet head with a single linear layer mapping the 512-dim pooled feature vector to our two classes. Training this network end-to-end gives it full flexibility to learn wildfire-specific filters, but it also demands careful optimization—high initial learning rates, aggressive augmentation, and adaptive schedulers—to avoid overfitting on our modest dataset. With appropriate regularization and a One-Cycle LR schedule, however, a scratch-trained ResNet-18 can still achieve strong performance by discovering features uniquely suited to flames and smoke.

Next, we leverage transfer learning by loading a ResNet-18 pre-trained on ImageNet, swapping out its final fully-connected layer for a new 512→2 head, and fine-tuning the entire network at a low learning rate. This approach benefits from generic visual representations—edges, textures, color gradients—already present in the backbone, allowing the model to converge faster and generalize better on limited wildfire data. By adjusting both the new head and the pretrained filters together, the network can refine its low-level detectors to pick up on subtle wildfire cues—faint smoke wisps, ember glows, or partially obscured flames—while still retaining the broad scene understanding learned from millions of natural images.

Finally, we experiment with a Vision Transformer (ViT-base-patch16_224) loaded with ImageNet weights. ViTs partition each image into 16×16 patches, embed them, and process them through transformer encoder blocks with multi-head self-attention. We replace the ViT's original classification head with a 768→2 linear layer and fine-tune at a similarly low learning rate. The global attention mechanism of ViT allows it to capture long-range relationships—useful for detecting smoke patterns that span wide areas—but its larger parameter count and weaker inductive biases can make it more sensitive to data scarcity. By comparing these three models, we assess the trade-off between the adaptability of scratch training, the efficiency of transfer learning, and the contextual power of transformer architectures.

4.4 Modifications

To better handle the class imbalance inherent in wildfire imagery datasets—where "no-fire" samples may vastly outnumber "fire" ones—we adopt a class-weighted loss strategy in the fine-tuning of our pretrained ResNet-18 model. Instead of altering the model's backbone structure or freezing policy, we calculate class

weights inversely proportional to their frequencies in the training data and integrate them into the CrossEntropyLoss function. This design aims to amplify the influence of underrepresented "fire" samples during gradient updates, nudging the model to prioritize subtle fire indicators that might otherwise be overlooked in favor of dominant background patterns. The overarching goal is to steer the classifier away from majority-class bias and toward a more balanced sensitivity across both categories—particularly vital in contexts where missing a fire detection carries high risk.

For the scratch-trained model, we introduce a more expressive classification head to increase the network's capacity for task-specific representation learning. Replacing the default linear layer, we employ a compact multi-layer perceptron consisting of a 512-unit bottleneck followed by BatchNorm1d, ReLU activation, Dropout with p=0.5, and a final projection to class logits. This architecture is designed to inject additional nonlinearity and regularization into the decision-making process, enabling the model to map raw convolutional features into a more discriminative embedding space. Batch normalization is included to stabilize the learning dynamics, while dropout is intended to mitigate overfitting—two factors that are particularly critical when training from scratch on a limited dataset.

Together, these architectural and training modifications seek to enhance the models' ability to generalize across complex wildfire imagery. By combining class-sensitive loss weighting with deeper classification heads, our configurations aim to cultivate models that are not only attentive to minority-class signals but also robust under real-world variability. These adjustments reflect a broader intention to tailor standard architectures for the unique demands of wildfire detection: ambiguous visual boundaries, sparse labeled data, and asymmetric risk profiles between false positives and false negatives.

5 Results - Comparative Analysis

The comparative evaluation of the scratch-trained ResNet18 models demonstrates that architectural and loss-based modifications can significantly enhance wildfire classification, even in the absence of pretrained weights. The MLP-head variant emerged as the strongest overall performer, achieving the highest test accuracy (82%) and the most balanced results across both classes, with F1-scores of 0.77 (fire) and 0.85 (no-fire)(Table 2). This suggests that incorporating nonlinearity and regularization into the classifier head enables the network to extract more nuanced and robust features—especially important in a domain with limited data and subtle visual distinctions.

By contrast, the linear-head model lagged notably behind, particularly on the fire class, where it scored the lowest across all three metrics: precision (0.69), recall (0.60), and F1-score (0.70). Despite slightly higher training accuracy, this variant failed to generalize effectively to the minority class, highlighting a tendency to overfit to dominant patterns present in the no-fire category. This clear performance gap underscores the value of even lightweight architectural improvements in facilitating better generalization for imbalanced classification tasks.

The model trained with a class-weighted loss function offered a meaningful intermediate solution. With a fire-class recall of 0.77 and an F1-score of 0.75, it significantly outperformed the linear-head model in minority class sensitivity, while maintaining a comparable overall accuracy (80%). Although it did not surpass the MLP-head variant in absolute terms, it demonstrated that loss rebalancing alone—without architectural changes—can provide measurable benefits when addressing class imbalance.

Taken together, these results validate the impact of targeted architectural and training modifications. The superior performance of the MLP-head and class-weighted models, particularly in contrast to the linear-head baseline, confirms that thoughtful configuration choices can yield competent, well-balanced wildfire detectors from scratch, even in the absence of transfer learning. These findings also emphasize that resolving trade-offs between precision and recall is key to optimizing classifier behavior in safety-critical, imbalanced detection tasks.

Fine-tuning a ResNet18 pretrained on ImageNet yielded a clear performance leap, with validation accuracy improving from 77–80% in the scratch models to 84–88%(Table 1). Precision and recall both increased across fire and no-fire classes, and F1-scores reached 0.85 (fire) and 0.90 (no-fire) on the initial dataset(Table 3). This improvement reflects the effectiveness of transferring learned representations—edges, textures, and semantic cues—from general image domains to the more specific context of wildfire detection. Importantly, even without synthetic data, the pretrained ResNet18 achieved strong balance between sensitivity and speci-

Model	**Train Loss**	**Train Accuracy**	**Validation Loss**	**Validation Accuracy**
ResNet18 (Scratch-MLP Head)	0.4371	0.8082	0.4669	0.7786
ResNet18 (Scratch-Linear Head)	0.4025	0.8232	0.5126	0.7711
ResNet18 (Scratch-Weighted Head)	0.4021	0.8193	0.4448	0.7861
ResNet18 (Fine-tuned)	0.3111	0.8691	0.3689	0.8408
ViT (Fine-tuned)	0.0920	0.9676	0.1718	0.9378

Table 1: Performance comparison on epoch 20.

Metric	Model					
	ResNet18 (Scratch-MLP Head)		ResNet18 (Scratch-Linear Head)		ResNet18 (Scratch-Weighted Head)	
	fire	nofire	fire	nofire	fire	nofire
Precision	0.76	0.86	0.69	0.88	0.72	0.85
Recall	0.79	0.84	0.60	0.93	0.77	0.81
F1 score	0.77	0.85	0.70	0.85	0.75	0.83
Accuracy	0.82		0.80		0.80	

Table 2: Classification Report Summary Across Non-Pretrained Models (Initial Dataset)

ficity, confirming the adaptability of its backbone features when modestly fine-tuned on a domain-relevant task.

Metric	Model			
	ResNet18 (Fine-tuned)		ViT (Fine-tuned)	
	fire	nofire	fire	nofire
Precision	0.82	0.92	0.94	0.94
Recall	0.89	0.88	0.91	0.96
F1 score	0.85	0.90	0.92	0.95
Accuracy	0.88		0.94	

Table 3: Classification Report Summary Across Pretrained Models (Initial Dataset)

The Vision Transformer (ViT), fine-tuned on the merged dataset enriched with synthetic images from the diffusion model, emerged as the best overall performer. It achieved a validation accuracy of 93.03% and an F1-score of 0.92 for fire and 0.95 for nofire class, exceeding all convolutional configurations in both precision and recall. ViT's ability to model long-range dependencies allowed it to distinguish nuanced fire features in cluttered or ambiguous scenes, such as partial smoke occlusions or distant flames in low contrast. These results suggest that ViT's attention-based structure may be especially well-suited for contexts like wildfire detection, where local texture cues alone are insufficient.

The evaluation of models on the merged dataset—including synthetic fire images generated via denoising diffusion—reveals that the anticipated performance gains did not materialize in practice (Tables 4 and 5). Contrary to expectations, no model demonstrated meaningful improvements in validation accuracy or class-wise performance metrics. The scratch-trained ResNet18 with an MLP head, for example, showed a marginal increase in training accuracy but yielded identical test F1-score and accuracy. Similarly, the fine-tuned ResNet18 exhibited a slight drop in validation accuracy (from 84.08% to 83.58%), and its fire-class precision and recall did not surpass previous results. Even the Vision Transformer, which consistently led in overall performance, showed a minor decline in validation accuracy (from 93.78% to 93.03%) and fire-class recall (from 0.93 to 0.91).

These results suggest that the addition of synthetic samples—despite increasing dataset size and diversity—did not directly enhance model generalization on the test set. One possible explanation is that the

synthetic fire images, while visually plausible, may not have introduced fundamentally new discriminative cues beyond those present in the real data. Another possibility is that domain shift or noise introduced by generative sampling diluted the signal quality for certain decision boundaries. Overall, these outcomes underscore that while synthetic data generation is a promising augmentation strategy, its practical impact depends heavily on the quality, variability, and task-specific alignment of the generated samples. Future work might explore more controlled generation methods, adversarial filtering, or curriculum-based integration to better exploit synthetic imagery in wildfire detection pipelines.

Model	**Train Loss**	**Train Accuracy**	**Validation Loss**	**Validation Accuracy**
ResNet18 (Scratch-MLP Head)	0.3809	0.8291	0.4119	0.8035
ResNet18 (Fine-tuned)	0.3050	0.8746	0.4313	0.8358
ViT (Fine-tuned)	0.0760	0.9733	0.1807	0.9303

Table 4: Performance comparison on merged dataset on epoch 20.

Metric	Model					
	ResNet18 (Scratch-MLP Head)		ResNet18 (Fine-tuned)		ViT (Fine-tuned)	
	fire	nofire	fire	nofire	fire	nofire
Precision	0.76	0.86	0.80	0.92	0.90	0.96
Recall	0.79	0.84	0.89	0.86	0.93	0.93
F1 score	0.77	0.85	0.84	0.89	0.91	0.94
Accuracy	0.82		0.87		0.93	

Table 5: Classification Report Summary Across Models (Merged Dataset)

6 Conclusion

In this work, we investigated the application of deep learning models for wildfire detection, evaluating a range of architectures and training strategies on a challenging, imbalanced visual dataset. Our results demonstrate that even in the absence of pretrained weights, thoughtful architectural and training modifications—such as adding nonlinearity through MLP heads and employing class-weighted loss—can substantially improve model performance, especially for underrepresented fire cases. These enhancements enabled our scratch-trained models to approach the performance of more resource-intensive pretrained configurations, offering a viable alternative in scenarios with limited access to large-scale pretrained models or computational resources.

Pretrained models, particularly the Vision Transformer, consistently delivered the best results, underscoring the value of transfer learning and advanced attention mechanisms in capturing subtle visual cues critical to fire detection. However, the marginal gains from synthetic data augmentation highlight the importance of aligning generative techniques with downstream task requirements.

Overall, our findings suggest that with well-designed model configurations and targeted training strategies, deep learning can support reliable wildfire detection across diverse environments. Such systems could be integrated into early warning platforms, drone-based surveillance, or environmental monitoring networks—enhancing our ability to respond quickly to emergent fire threats and ultimately contributing to more resilient, data-driven disaster management infrastructures.

References

[1] Amila Akagic and Emir Buza. Lw-fire: A lightweight wildfire image classification with a deep convolutional neural network. *Applied Sciences*, 12:2646, 03 2022.

[2] Ismail El-Madafri, Marta Peña, and Noelia Olmedo-Torre. The wildfire dataset: Enhancing deep learning-based forest fire detection with a diverse evolving open-source dataset focused on data representativeness and a novel multi-task learning approach. *Forests*, 14(9), 2023.

[3] Yassine El-Madafri. The wildfire dataset. https://www.kaggle.com/datasets/elmadafri/the-wildfire-dataset, 2022. Accessed: 2025-06-15.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[5] Javier Selva, Anders S. Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B. Moeslund, and Albert Clapés. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12922–12943, November 2023.

[6] Gowtham Raj Vuppari, Navarun Gupta, Ahmed El-Sayed, and Xingguo Xiong. Wildfire detection using vision transformer with the wildfire dataset, 2025.