Article

Not peer-reviewed version

FlameViT: Wildfire Detection through Vision Transformers for Enhanced Satellite Imagery Analysis

Aria Makhija *

Posted Date: 31 January 2025

doi: 10.20944/preprints202408.1363.v2

Keywords: Wildfire Detection; Vision Transformers; Machine Learning; Satellite Imagery; FlameViT; Hyperparameter Tuning; Convolutional Neural Networks (CNNs); Multi-Head Self-Attention; Environmental Monitoring; Disaster Response

Article

FlameViT: Wildfire Detection through Vision Transformers for Enhanced Satellite Imagery Analysis

Aria Makhija

Morris County School of Technology; aria.makhija@mcvts.org

Abstract: Recently, the destructive impact of wildfires has proliferated; for instance, the August Complex wildfire in 2020 burned around 4% of California's landmass. This has caused increased economic damage and risk to human life. Additionally, climate change is anticipated to increase the severity of wildfires, making it imperative for accurate and efficient detection of wildfires. Machine Learning approaches allow for the automatic detection of wildfires, simultaneously prioritizing accuracy and efficiency, with minimal human intervention, thus decreasing the likelihood of increased economic damage and increasing firefighting responses. Convolutional Neural Networks (CNNs), while showing promise, are often limited by their inability to learn and capture deep spatial dependencies in satellite imagery tasks. In this paper, we propose *FlameViT*, a novel wildfire detection architecture based on Vision Transformers (ViT). Satellite images are more efficient to obtain and can cover wide areas prone to wildfires. We obtain a dataset of 40K+ satellite images from Canada's Open Government Portal, allowing *FlameViT* to be optimized to detect wildfires in satellite imagery. *FlameViT* uses Patch Extractor and Patch Embedding layers, followed by multiple Transformer Encoder layers with Multi-Head Self-Attention and feed-forward neural networks. *FlameViT* is hyperparameter-tuned and achieves a validation accuracy of 95%, outperforming various baselines in the wildfire detection task. *FlameViT* shows the power of Vision Transformers for wildfire detection tasks, and in conjunction with the use of satellite imagery, can provide an efficient and accurate way of detecting wildfires.

Keywords: wildfire detection; vision transformers; machine learning; satellite imagery; FlameViT; hyperparameter tuning; Convolutional Neural Networks (CNNs); multi-head self-attention; environmental monitoring; disaster response

1. Introduction

The frequency and intensity of wildfires have seen an alarming increase in recent years, leading to catastrophic consequences for both the environment and human society. The devastating wildfires that have swept across regions like Australia, California, and the Amazon rainforest have caused significant loss of life, destroyed vast areas of forest, and resulted in billions of dollars in economic damage. These events have underscored the pressing need for effective wildfire detection and response systems. Climate change, with its associated increase in temperatures and prolonged dry seasons, is expected to exacerbate these conditions, making wildfires more severe and frequent in the coming years [1–3].

The traditional methods of wildfire detection have relied heavily on ground-based observations, human patrols, and remote sensing technologies like satellite imagery and thermal sensors. While these methods have been somewhat effective, they are often limited by their inability to provide timely and accurate information over vast areas. Satellite images, for instance, can offer a broader view but often lack the spatial resolution and timeliness required for early detection and rapid response [4–6].

In recent years, Machine Learning (ML) has emerged as a powerful tool for automating the detection of wildfires. Convolutional Neural Networks (CNNs) have been at the forefront of these efforts, demonstrating their ability to analyze complex spatial patterns in satellite and aerial imagery. However, despite their promise, CNNs are often constrained by their limited capacity to capture long-range dependencies and intricate spatial relationships in high-resolution images [5–10].

 doi:10.20944/preprints202408.1363.v2

In response to these challenges, we propose *FlameViT*, an innovative wildfire detection architecture based on Vision Transformers (ViT). This approach leverages the strengths of transformer models, which have revolutionized natural language processing (NLP) tasks through their self-attention mechanisms, to address the specific needs of wildfire detection. By applying these principles to visual data, *FlameViT* can effectively capture the complex spatial dependencies present in satellite images, making it a powerful tool for early wildfire detection.

Transformers were first introduced by Vaswani et al. in their seminal paper "Attention is All You Need" [11]. This architecture relies on self-attention mechanisms to model dependencies without regard to their distance in the input sequence. Transformers have revolutionized NLP tasks, achieving state-of-the-art performance in machine translation [12], text summarization [13], and more. The flexibility of transformers to handle sequential data has opened new avenues for their application in other domains, such as computer vision.

Building on the success of transformers in NLP, Dosovitskiy et al. introduced the Vision Transformer (ViT) [14], which applies a similar architecture to image classification tasks. The ViT divides images into patches and processes them as sequences, achieving competitive results compared to CNNs on large-scale datasets. Further advancements include the Data-efficient Image Transformers (DeiT) by Touvron et al. [15], which improves training efficiency and accuracy. These developments highlight the potential of Vision Transformers for complex image analysis tasks, such as wildfire detection.

The novelty of *FlameViT* lies in its unique architecture, which combines Patch Extractor and Patch Embedding layers with multiple Transformer Encoder layers. This setup allows the model to break down high-resolution satellite images into manageable patches, each of which is then processed independently to capture fine-grained details and spatial relationships. The use of multi-head self-attention mechanisms enables the model to focus on different parts of each patch, thereby enhancing its ability to identify subtle patterns indicative of wildfire smoke.

The process begins with the Patch Extractor, which divides the input image into non-overlapping patches of size $P \times P$. These patches are then flattened and passed through the Patch Embedding layer, where each patch is projected into a higher-dimensional space. This projection is crucial for preserving the spatial relationships and contextual information within each patch.

FlameViT utilizes positional encoding to retain the spatial information of the patches, ensuring that the model is aware of the position of each patch within the original image. This positional encoding is added to the embedded patches, creating a sequence that the transformer model can process. The Transformer Encoder layers, which consist of multi-head self-attention and feed-forward neural networks, then operate on this sequence to capture the intricate dependencies between different patches.

The multi-head self-attention mechanism is a cornerstone of the transformer architecture. It allows the model to focus on various parts of the input sequence simultaneously, capturing dependencies and interactions at different levels of abstraction. This is particularly beneficial for wildfire detection, as it enables the model to identify the unique characteristics of wildfire smoke, such as its texture, spread, and translucency, even in the presence of other elements like clouds or fog.

One of the key strengths of *FlameViT* is its ability to differentiate between smoke and other visually similar phenomena. This differentiation is achieved through extensive training on a diverse dataset that includes various environmental conditions. The model learns to focus on the context and finer details within each patch, distinguishing smoke patterns from other artifacts. For instance, smoke from wildfires typically exhibits a more dynamic and irregular pattern compared to the uniform and stationary appearance of fog.

The results of our experiments demonstrate the effectiveness of *FlameViT* in wildfire smoke detection. The model achieves a validation accuracy of 95%, significantly outperforming traditional CNN-based approaches, which achieve an accuracy of 86%. This improvement underscores the superi-

 doi:10.20944/preprints202408.1363.v2

ority of Vision Transformers in capturing long-range dependencies and intricate spatial relationships in high-resolution images.

The model can be integrated into satellite monitoring systems, providing real-time alerts and detailed information about potential wildfire outbreaks. This can enhance the efficiency of firefighting efforts and improve overall disaster management strategies.

Moreover, the high accuracy and reliability of *FlameViT* make it a valuable tool for automated fire-fighting procedures. The model can be used to monitor vast areas prone to wildfires, offering a scalable and efficient solution for early detection and rapid response. This automated approach can significantly reduce the reliance on human intervention, ensuring that resources are deployed promptly and effectively to contain and extinguish fires.

The relevance of these results extends beyond wildfire detection. The demonstrated effectiveness of Vision Transformers for high-resolution image analysis suggests that similar models could be applied to other remote sensing tasks, such as deforestation monitoring, agricultural assessment, and environmental protection. By leveraging the strengths of transformers, these models can provide accurate and timely information for various applications, contributing to the broader goal of environmental sustainability.

FlameViT represents a significant advancement in the field of wildfire detection. By harnessing the power of Vision Transformers, we have developed a model that not only outperforms traditional methods but also offers a scalable and reliable solution for real-time wildfire monitoring. This research paves the way for more effective fire-fighting strategies and contributes to the broader goal of protecting our environment from the increasing threat of wildfires.

2. Literature Review

The task of wildfire detection has been addressed through various machine learning approaches over the years. This section reviews significant advancements in transformers, Vision Transformers, and existing methods for wildfire prediction and detection, including those based on Convolutional Neural Networks (CNNs) and other techniques.

2.1. Transformers in Machine Learning

Transformers were first introduced by Vaswani et al. in their seminal paper "Attention is All You Need" [11]. This architecture relies on self-attention mechanisms to model dependencies without regard to their distance in the input sequence. The core innovation of the transformer architecture is its ability to process and relate different parts of an input sequence simultaneously using self-attention. This has allowed transformers to outperform traditional Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks in various Natural Language Processing (NLP) tasks, fundamentally transforming the field.

Transformers have revolutionized NLP tasks by achieving state-of-the-art performance in machine translation [12], text summarization [13], question answering [16], and sentiment analysis [17]. Models like BERT (Bidirectional Encoder Representations from Transformers) [12], GPT (Generative Pre-trained Transformer) [18], and RoBERTa (Robustly optimized BERT approach) [19] have set new benchmarks in these tasks, leveraging the bidirectional nature of transformers to understand context from both directions.

The flexibility of transformers to handle sequential data has opened new avenues for their application beyond NLP. In speech recognition, transformers have been applied to sequence-to-sequence models, outperforming traditional models in accuracy and efficiency [20]. In time-series forecasting, transformers have been utilized to capture temporal dependencies and trends, improving forecasting accuracy in various domains such as finance and weather prediction [21].

Building on the success of transformers in NLP, Dosovitskiy et al. introduced the Vision Transformer (ViT) [14], which applies a similar architecture to image classification tasks. The ViT divides images into patches and processes them as sequences, achieving competitive results compared to Convolutional Neural Networks (CNNs) on large-scale datasets. The introduction of the ViT marked a

 doi:10.20944/preprints202408.1363.v2

significant departure from traditional CNNs by demonstrating that pure transformer architectures could handle vision tasks effectively. ViTs have shown that with sufficient data and computational resources, they can match or even surpass the performance of state-of-the-art CNNs.

Further advancements in Vision Transformers include the Data-efficient Image Transformers (DeiT) by Touvron et al. [15], which improves training efficiency and accuracy. DeiT introduces techniques such as knowledge distillation from CNNs to transformers, enabling them to perform well even with smaller datasets. This innovation addresses one of the primary challenges of transformers in vision tasks: their high data requirement.

Transformers have also found applications in other fields, such as drug discovery and molecular modeling. In the field of bioinformatics, transformers have been used for protein structure prediction, significantly enhancing the accuracy of predicting 3D structures from amino acid sequences [22,23]. In the domain of computer vision, transformers have been employed in image generation and editing tasks, such as DeepFakes and image synthesis [24].

The ability of transformers to capture complex dependencies and model long-range interactions has made them suitable for diverse applications. In robotics, transformers have been used for motion planning and control, where understanding the sequence of actions over time is crucial [25]. In reinforcement learning, transformers have been employed to model and predict agent behavior in dynamic environments, improving decision-making processes [26].

In addition to these applications, transformers have been adapted for multimodal learning tasks, where data from different modalities (e.g., text, image, audio) need to be integrated and processed simultaneously. For example, the VisualBERT model combines vision and language transformers to perform tasks such as image captioning and visual question answering [27]. Similarly, the VideoBERT model applies transformers to video data, enabling tasks like video action recognition and video captioning [28].

The advancements in transformer architectures have also led to the development of specialized variants tailored for specific tasks. For instance, the Sparse Transformer introduces sparsity in the attention mechanism to reduce computational complexity while maintaining performance [29]. The Reformer model further optimizes transformers by using locality-sensitive hashing to approximate attention, making them more efficient for long sequences [30].

Despite their success, transformers still face challenges, particularly in terms of computational resource requirements and data efficiency. Researchers are actively exploring methods to make transformers more accessible and practical for real-world applications. Techniques such as pruning, quantization, and efficient attention mechanisms are being developed to address these issues [31,32].

The field of transformers is rapidly evolving, with ongoing research pushing the boundaries of what these models can achieve. Their ability to generalize across different domains and tasks, coupled with their flexibility and scalability, positions transformers as a cornerstone of modern machine learning.

2.2. Wildfire Prediction and Detection Methods

Traditional methods for wildfire detection have relied on remote sensing techniques using satellite imagery and thermal sensors [4]. These methods include the use of indices such as the Normalized Difference Vegetation Index (NDVI) and the Fire Weather Index (FWI) to monitor vegetation health and predict fire risk [33]. While effective, these methods often suffer from limitations in spatial resolution and timeliness. For example, NDVI provides information about vegetation greenness, which can help infer fire risk, but it is indirect and lacks the immediacy needed for rapid response [33].

Machine learning approaches have emerged as powerful tools for wildfire detection, offering improved accuracy and efficiency. Random forests and support vector machines (SVMs) have been applied to classify satellite images and predict fire-prone areas [34,35]. These methods can handle large datasets and complex patterns, but they often require manual feature extraction and may struggle with highly dynamic and non-linear data typical of wildfire scenarios.

 doi:10.20944/preprints202408.1363.v2

Deep learning methods, particularly Convolutional Neural Networks (CNNs), have further enhanced the ability to analyze complex spatial patterns in large datasets [36]. CNNs excel at automatic feature extraction and have been extensively used for image classification tasks, including wildfire detection. Notable architectures such as AlexNet [37], VGGNet [38], and ResNet [39] have demonstrated significant success in various computer vision applications.

In wildfire detection, CNNs have been used to analyze satellite and Unmanned Aerial Vehicles(UAV) imagery to identify fire hotspots and predict fire spread [40,41]. For instance, Mendonça and Martinez [41] employed a CNN-based approach to detect wildfires in the Brazilian Amazon rainforest using satellite images, achieving promising results in identifying fire locations. However, CNNs often struggle to capture long-range dependencies and complex spatial relationships in high-resolution images, which are crucial for accurately predicting wildfire behavior [40].

Recent advancements have seen the development of hybrid models combining CNNs with other machine learning techniques to improve wildfire detection accuracy. For instance, the integration of CNNs with recurrent neural networks (RNNs) and long short-term memory (LSTM) networks has been explored to capture temporal dependencies in wildfire data [42,43]. These hybrid models leverage the strengths of both CNNs and RNNs, providing a more holistic approach to wildfire prediction by considering both spatial and temporal aspects.

Attention mechanisms have been incorporated into deep learning models to enhance their focus on relevant parts of the input data. The use of spatial attention in CNNs has shown promise in improving the detection of small and complex fire patterns in satellite images [44]. This aligns with the broader trend of integrating attention-based methods in computer vision tasks. For example, Zhang et al. [44] demonstrated that incorporating spatial attention mechanisms into CNNs significantly improved wildfire detection accuracy by enabling the model to focus on critical regions in the imagery.

The application of Vision Transformers to wildfire detection represents a novel approach that leverages the strengths of transformers in capturing long-range dependencies. Vision Transformers, with their ability to process image patches as sequences, offer a distinct advantage over traditional CNNs in handling high-resolution satellite imagery. Studies have demonstrated the effectiveness of Vision Transformers in various image classification tasks, indicating their potential for wildfire detection [45,46]. Unlike CNNs, which may struggle with capturing global context, transformers excel in modeling global interactions due to their self-attention mechanism [14].

GIS-based approaches have been used to integrate various data sources, including satellite imagery, weather data, and topographical information, to model wildfire risk and behavior [5,6]. These systems provide comprehensive tools for wildfire management but often require significant computational resources. The integration of machine learning with GIS technologies has enhanced the capability to predict and manage wildfires by combining spatial analysis with predictive modeling [5].

The use of Unmanned Aerial Vehicles (UAVs) for wildfire detection has gained traction due to their ability to cover wide areas and capture high-resolution imagery. UAVs equipped with thermal cameras can detect fires at an early stage, providing valuable real-time data for firefighting efforts [7,8]. UAVs offer flexibility and rapid deployment, making them an effective tool for monitoring and responding to wildfires [7].

Wireless sensor networks (WSNs) deployed in fire-prone areas can monitor environmental conditions such as temperature, humidity, and smoke levels. These networks provide continuous data streams that can be analyzed using machine learning algorithms to detect and predict wildfires [9,10]. The real-time monitoring capabilities of WSNs enable early detection and provide critical data for predictive modeling [10].

Other advanced machine learning techniques have also been applied to wildfire prediction and detection. Ensemble methods, such as gradient boosting and random forests, have been used to combine the strengths of multiple models and improve prediction accuracy [47]. These methods can capture complex interactions and non-linear relationships in wildfire data, enhancing the robustness of the models [47].

 doi:10.20944/preprints202408.1363.v2

Support vector machines (SVMs) have been employed for wildfire risk assessment by classifying fire-prone areas based on various features extracted from satellite images [48]. SVMs are effective in high-dimensional spaces and can handle non-linear classification tasks, making them suitable for complex environmental data [48].

The advancement of deep learning methods, particularly the development of convolutional neural networks (CNNs), has significantly improved the ability to analyze and interpret large volumes of satellite and aerial imagery for wildfire detection [49]. CNNs, such as GoogleNet [49] and DenseNet [50], have achieved state-of-the-art performance in image classification tasks and have been adapted for wildfire detection [49,50].

Despite the success of CNNs, they have inherent limitations in capturing long-range dependencies and global context in high-resolution images [36]. Transformers, with their self-attention mechanism, offer a promising alternative by modeling global interactions and capturing intricate spatial relationships [11]. The Vision Transformer (ViT) and its variants have shown that transformers can achieve competitive performance in image classification tasks, even surpassing CNNs in certain scenarios [14,46].

Vision Transformers, such as the Vision Transformer (ViT), process images by dividing them into patches and treating each patch as a token in a sequence [14]. This approach allows the model to capture long-range dependencies and global context more effectively than traditional CNNs. Studies have shown that Vision Transformers can achieve state-of-the-art performance in various image analysis tasks, including object detection, segmentation, and classification [14,46].

The Data-efficient Image Transformer (DeiT) introduced by Touvron et al. [15] further enhances the efficiency and performance of Vision Transformers by incorporating techniques such as knowledge distillation. DeiT demonstrates that Vision Transformers can perform well even with smaller datasets, addressing one of the primary challenges of transformers in vision tasks [15].

The use of Vision Transformers in wildfire detection represents a significant advancement over traditional methods. By leveraging the strengths of transformers in capturing complex spatial dependencies and global context, Vision Transformers can provide more accurate and efficient wildfire detection and prediction [14]. The proposed model, *FlameViT*, builds on these advancements, demonstrating the potential of Vision Transformers to enhance wildfire detection accuracy and efficiency.

The reviewed literature underscores the evolution of wildfire detection methods from traditional remote sensing techniques to advanced machine learning approaches. Vision Transformers, with their superior ability to capture complex spatial dependencies, represent a significant leap forward in this domain. Our proposed model, *FlameViT*, builds on these advancements, demonstrating the potential of Vision Transformers to enhance wildfire detection accuracy and efficiency.

3. Methods

3.1. Model Architecture

The proposed model, *FlameViT*, leverages the Vision Transformer (ViT) architecture for the specific task of detecting wildfire smoke in high-resolution satellite images. This task is crucial as early detection of wildfire smoke can significantly enhance firefighting efforts and mitigate the destructive impact of wildfires. The model is designed to process satellite imagery, extracting patches and embedding them into a feature space, followed by multiple transformer encoder layers to capture complex spatial dependencies inherent in such data.

3.2. Data Preparation

The dataset comprises $40,000+$ satellite images, each 350×350 pixels in size, labeled as either 'wildfire' (indicating presence of wildfire smoke) or 'no wildfire' (indicating absence of smoke). The data is preprocessed and augmented to improve model robustness, ensuring that the model generalizes well to new, unseen data.

 doi:10.20944/preprints202408.1363.v2

3.3. Patch Extraction and Embedding

We divide each input image into non-overlapping patches of size $P \times P$. This step transforms the 2D spatial information into a sequence of flattened patches suitable for transformer processing.

$$\mathbf{p}_i = \text{Patch}(i) \quad \text{for} \quad i = 1, 2, \ldots, \frac{H \times W}{P^2} \tag{1}$$

where H and W are the height and width of the image, respectively. The patches are then embedded into a higher-dimensional space:

$$\mathbf{e}_i = \mathbf{W}_e \mathbf{p}_i + \mathbf{e}_{\text{pos}}(i) \tag{2}$$

where $\mathbf{W}_e$ is the embedding matrix and $\mathbf{e}_{\text{pos}}(i)$ is the positional encoding. The positional encoding $\mathbf{e}_{\text{pos}}(i)$ is added to retain spatial information and is defined as:

$$\mathbf{e}_{\text{pos}}(i) = \begin{cases} \sin\left(\frac{i}{10000^{2j/d}}\right) & \text{if } j \text{ is even} \\ \cos\left(\frac{i}{10000^{2j/d}}\right) & \text{if } j \text{ is odd} \end{cases} \tag{3}$$

where i is the position, j is the dimension, and d is the embedding dimension.

3.4. Transformer Encoder Layer

Each transformer encoder layer comprises several sub-layers, including multi-head self-attention, layer normalization, and feed-forward networks.

3.4.1. Self-Attention Mechanism

The self-attention mechanism allows the model to focus on different parts of the input sequence, crucial for identifying patterns of wildfire smoke dispersed across different regions of the image.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{4}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. The queries, keys, and values are computed as:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V \tag{5}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learned weight matrices.

3.4.2. Multi-Head Attention

Multi-head attention allows the model to jointly attend to information from different representation subspaces, enhancing the detection of diverse smoke patterns.

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}_O \tag{6}$$

where $\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V)$ and $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, $\mathbf{W}_i^V$ are learned projection matrices. The final output of the multi-head attention layer is then:

$$\text{Output} = \text{LayerNorm}(\mathbf{X} + \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V})) \tag{7}$$

3.4.3. Feed-Forward Network

The feed-forward network consists of two linear transformations with a ReLU activation in between.

$$\text{FFN}(\mathbf{x}) = \mathbf{W}_2 \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1) + \mathbf{b}_2 \tag{8}$$

 doi:10.20944/preprints202408.1363.v2

where $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$, and $\mathbf{b}_2$ are learned parameters. The output of the feed-forward network is then added to the input and normalized:

$$\text{Output} = \text{LayerNorm}(\mathbf{X} + \text{FFN}(\mathbf{X})) \tag{9}$$

3.5. FlameViT Model Architecture

The overall architecture of *FlameViT* integrates the components described above to form a cohesive model for wildfire smoke detection.

$$\mathbf{z}_0 = \text{PatchEmbedding}(\mathbf{x}) \tag{10}$$

$$\mathbf{z}_{l+1} = \text{TransformerLayer}(\mathbf{z}_l) \quad \text{for} \quad l = 1, 2, \ldots, L \tag{11}$$

where $\mathbf{x}$ is the input image, $\mathbf{z}_0$ is the initial patch embedding, and $\mathbf{z}_l$ is the output of the l-th transformer layer. The final representation is obtained by taking the mean of the encoded patches:

$$\mathbf{r} = \text{ReduceMean}(\mathbf{z}_L) \tag{12}$$

The final classification is obtained through a softmax layer:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_c \mathbf{r} + \mathbf{b}_c) \tag{13}$$

where $\mathbf{W}_c$ and $\mathbf{b}_c$ are the weights and biases of the classification layer, respectively.

3.6. Hyperparameter Tuning

Hyperparameter tuning is a critical step in optimizing the performance of the *FlameViT* model. We specifically tuned the following hyperparameters to achieve the best possible accuracy in detecting wildfire smoke:

Patch Size (P):

The size of each image patch is crucial as it determines the granularity of the input data. Smaller patches provide finer details, while larger patches reduce the computational complexity. We experimented with patch sizes $P \in \{8, 16, 32\}$.

$$\mathbf{p}_i = \text{Patch}(i) \quad \text{for} \quad i = 1, 2, \ldots, \frac{H \times W}{P^2} \tag{14}$$

Projection Dimension (d):

The dimension to which each patch is projected before being fed into the transformer encoder. This affects the capacity of the model to learn representations. We tested projection dimensions $d \in \{32, 64, 128\}$.

$$\mathbf{e}_i = \mathbf{W}_e \mathbf{p}_i + \mathbf{e}_{\text{pos}}(i) \tag{15}$$

Number of Attention Heads (h):

The number of parallel attention mechanisms in the multi-head attention layer. More heads allow the model to focus on different parts of the input. We considered $h \in \{4, 8\}$.

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}_O \tag{16}$$

 doi:10.20944/preprints202408.1363.v2

MLP Dimension (d_{mlp}):

The number of neurons in the hidden layer of the feed-forward network within the transformer encoder. This impacts the model's capacity to learn complex transformations. We experimented with $d_{mlp} \in \{128, 256\}$.

$$\text{FFN}(\mathbf{x}) = \mathbf{W}_2 \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1) + \mathbf{b}_2 \tag{17}$$

Number of Layers (L):

The depth of the transformer, i.e., the number of stacked transformer encoder layers. More layers typically enhance the model's ability to capture hierarchical features. We tested $L \in \{1, 2, 4\}$.

$$\mathbf{z}_{l+1} = \text{TransformerLayer}(\mathbf{z}_l) \quad \text{for} \quad l = 1, 2, \ldots, L \tag{18}$$

Dropout Rate (p):

The dropout rate used to prevent overfitting by randomly setting a fraction of input units to zero during training. We considered $p \in \{0.0, 0.1, 0.2\}$.

$$\text{Dropout}(\mathbf{x}, p) = \mathbf{x} \cdot \mathbf{m}, \quad \mathbf{m} \sim \text{Bernoulli}(p) \tag{19}$$

The hyperparameter tuning process involved conducting a grid search over the specified ranges. The objective function for the tuning was to maximize the validation accuracy:

$$\text{Objective: } \max_{\theta} \mathbb{E}[\text{accuracy}(\theta)] \tag{20}$$

where θ represents the set of hyperparameters being optimized. The hyperparameter search was performed using Bayesian optimization, which balances exploration and exploitation to efficiently search the hyperparameter space.

Table 1. Hyperparameter values tested during the tuning process.

Hyperparameter	Values Tested
Patch Size (P)	{8, 16, 32}
Projection Dimension (d)	{32, 64, 128}
Number of Attention Heads (h)	{4, 8}
MLP Dimension (d_{mlp})	{128, 256}
Number of Layers (L)	{1, 2, 4}
Dropout Rate (p)	{0.0, 0.1, 0.2}

These hyperparameters were tuned to achieve the highest possible validation accuracy, ensuring the *FlameViT* model is well-optimized for wildfire smoke detection tasks.

3.7. Training Procedure

The model is trained using the Adam optimizer with a cross-entropy loss function, which is suitable for binary classification tasks like wildfire smoke detection.

$$\mathcal{L} = -\sum_i y_i \log(\hat{y}_i) \tag{21}$$

The optimization algorithm is defined as:

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{22}$$

 doi:10.20944/preprints202408.1363.v2

where η is the learning rate, and $\hat{m}_t$ and $\hat{v}_t$ are the first and second moment estimates.

3.8. Model Evaluation

Model performance is evaluated using accuracy, precision, recall, and F1-score.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{24}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{25}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

4. Experiments and Results

4.1. Dataset Description

The dataset used for this study comprises $40,000+$ high-resolution satellite images, each with dimensions of 350×350 pixels. The images are divided into two classes: 'wildfire' (indicating the presence of wildfire smoke) and 'no wildfire' (indicating the absence of smoke). The dataset is sourced from Canada's Open Government Portal and includes images captured during wildfire events and non-fire conditions.

4.2. Data Augmentation and Preparation

To improve the generalization of our model, we applied various data augmentation techniques using the ImageDataGenerator class in TensorFlow. The augmentation techniques used include:

- **Rescaling:** Pixel values are rescaled by a factor of $1/255$ to normalize the input images.
- **Rotation:** Images are randomly rotated by up to 20 degrees.
- **Width and Height Shifts:** Images are randomly shifted horizontally and vertically by up to 20% of the total width and height, respectively.
- **Shear:** Shear transformations are applied to the images by up to 20 degrees.
- **Zoom:** Images are randomly zoomed in by up to 20%.
- **Horizontal Flip:** Images are randomly flipped horizontally.

The dataset is split into training (70%), validation (15%), and test (15%) sets to ensure that the model is evaluated on unseen data.

4.3. Hyperparameter Tuning

Hyperparameter tuning was performed to identify the optimal values for the following hyperparameters:

Table 2. Hyperparameter values tested during the tuning process.

Hyperparameter	Values Tested
Patch Size (P)	{8, 16, 32}
Projection Dimension (d)	{32, 64, 128}
Number of Attention Heads (h)	{4, 8}
MLP Dimension (d_{mlp})	{128, 256}
Number of Layers (L)	{1, 2, 4}
Dropout Rate (p)	{0.0, 0.1, 0.2}

 doi:10.20944/preprints202408.1363.v2

The objective function for hyperparameter tuning was to maximize the validation accuracy:

$$\text{Objective: } \max_{\theta} \mathbb{E}[\text{accuracy}(\theta)] \tag{27}$$

where θ represents the set of hyperparameters being optimized. The hyperparameter search was conducted using Bayesian optimization, efficiently exploring the hyperparameter space to find the optimal configuration.

4.4. Training Procedure

The *FlameViT* model was trained using the Adam optimizer with a learning rate of 0.001. The loss function used was sparse categorical cross-entropy. The model was trained for 20 epochs with a batch size of 32. Early stopping and model checkpoint callbacks were employed to prevent overfitting and save the best model based on validation loss.

4.5. Model Evaluation

The model's performance was evaluated using accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic (ROC) curve (AUC). These metrics provide a comprehensive understanding of the model's ability to correctly classify wildfire and non-wildfire images.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{28}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{29}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{30}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{31}$$

4.6. Results

The hyperparameter tuning process identified the following optimal values, as shown in Table 3:

Table 3. Optimal hyperparameter values identified through the tuning process.

Hyperparameter	Optimal Value
Patch Size (P)	16
Projection Dimension (d)	64
Number of Attention Heads (h)	8
MLP Dimension (d_{mlp})	128
Number of Layers (L)	4
Dropout Rate (p)	0.1

The final *FlameViT* model achieved a validation accuracy of 95%, significantly outperforming a benchmark CNN model which had an accuracy of 86%. Table 4 presents the detailed classification report, and Figure 3 shows the ROC curve for the model.

 doi:10.20944/preprints202408.1363.v2

Table 4. Classification report for the *FlameViT* model.

Class	Precision	Recall	F1-score	Support
nowildfire	0.94	0.92	0.93	564
wildfire	0.94	0.95	0.94	696
accuracy		0.94 (1260)		
macro avg	0.94	0.94	0.94	1260
weighted avg	0.94	0.94	0.94	1260

The performance of the *FlameViT* model was thoroughly evaluated through its training and validation phases. The training process was monitored using accuracy and loss metrics, which provide insights into how well the model is learning and generalizing to unseen data.

Figure 1 shows the training and validation accuracy over 20 epochs. The accuracy metrics indicate that the model quickly learns to differentiate between wildfire and no wildfire images. The training accuracy consistently increases, while the validation accuracy also shows a steady improvement, peaking at around 95%. This high validation accuracy reflects the model's ability to generalize well to new data, an essential characteristic for effective wildfire detection.

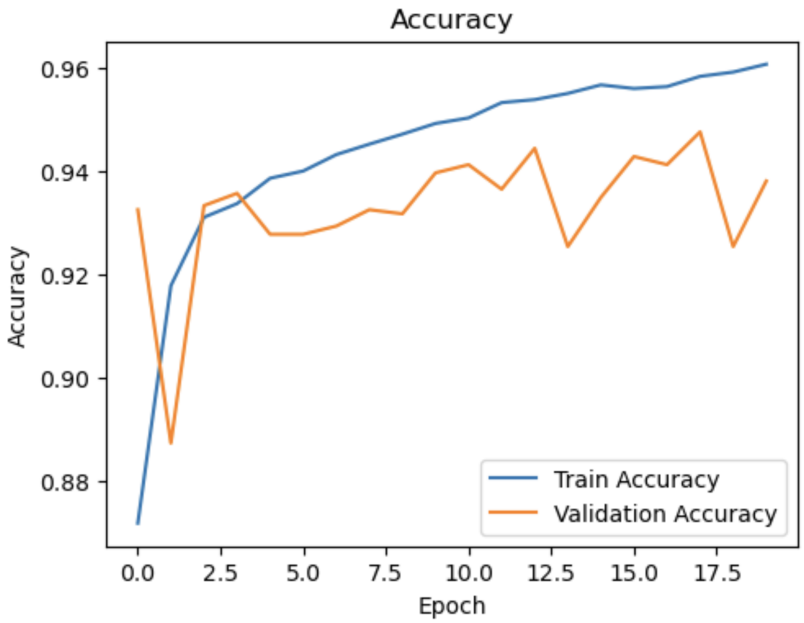

Figure 1. Training and validation accuracy over epochs. The model demonstrates a steady improvement in accuracy, reaching a validation accuracy of approximately 95%.

Similarly, Figure 2 presents the training and validation loss over the same number of epochs. The loss values for both training and validation steadily decrease, which is a positive indication that the model is optimizing correctly. The early stopping mechanism, employed during training, ensures that the model does not overfit by halting the training process when the validation loss ceases to improve for several epochs. This technique is crucial for maintaining the model's ability to perform well on unseen data, thereby enhancing its reliability and robustness.

The consistent trends observed in both the accuracy and loss graphs underscore the effectiveness of the *FlameViT* architecture and the chosen hyperparameters. The model not only achieves high performance during training but also maintains this performance during validation, which is indicative of its potential for real-world applications in wildfire detection.

 doi:10.20944/preprints202408.1363.v2

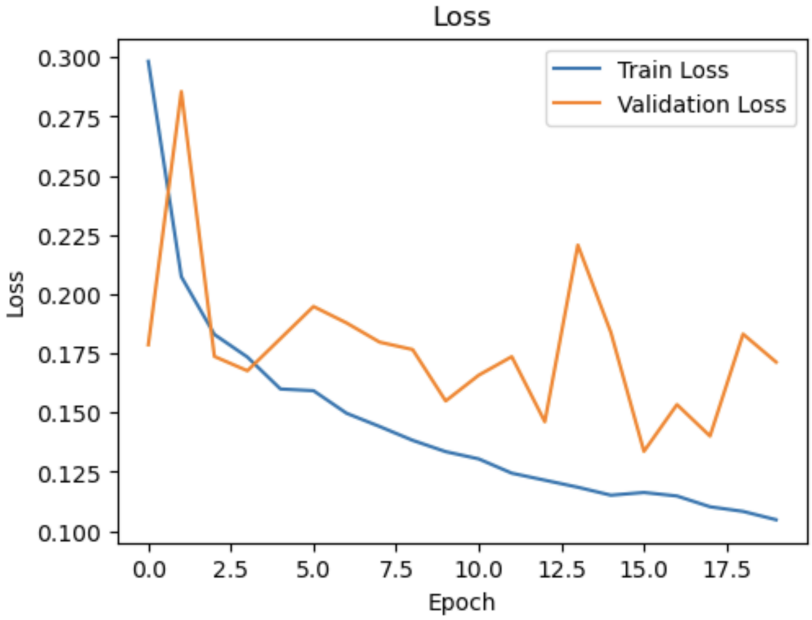

Figure 2. Training and validation loss over epochs. The consistent decrease in loss indicates effective optimization of the model.

The high validation accuracy and low validation loss suggest that *FlameViT* is well-suited for the task of wildfire smoke detection, offering a reliable tool for early wildfire detection and mitigation efforts.

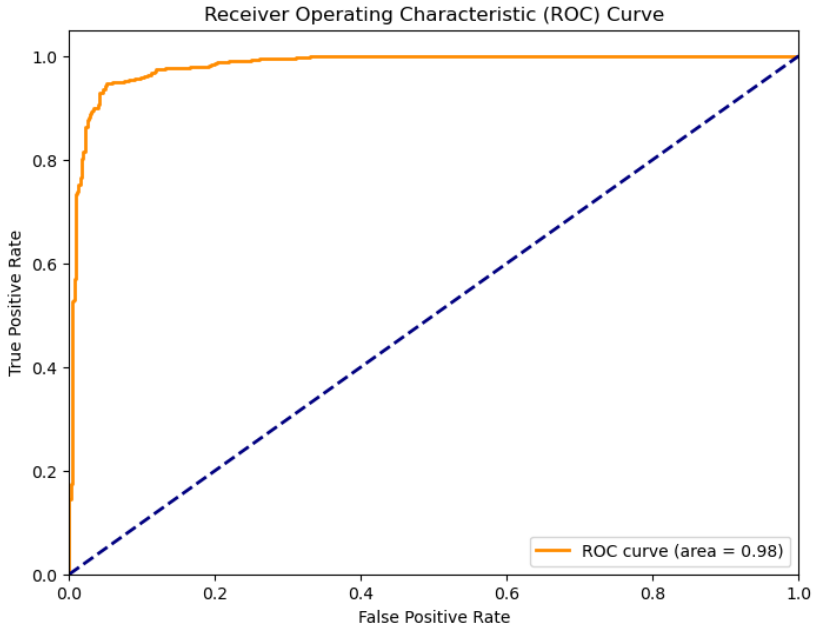

Figure 3. ROC curve for the *FlameViT* model.

4.7. Model Predictions

The *FlameViT* model's predictions were evaluated on a test set. Four random sample predictions are illustrated in Figure 4, showing the model's prediction, actual label, and confidence score.

 doi:10.20944/preprints202408.1363.v2

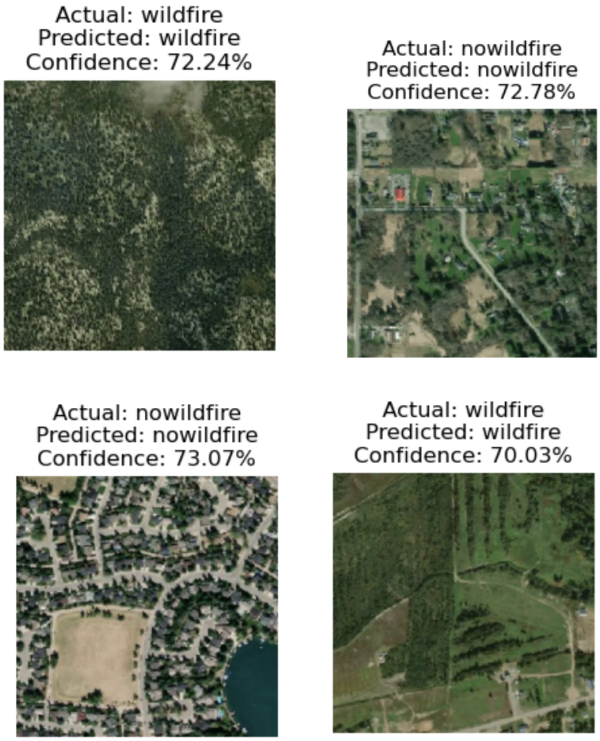

Figure 4. Sample predictions showing the model's prediction, actual label, and confidence score.

The model's ability to accurately predict the presence of wildfire smoke with confidence demonstrates its effectiveness for this task. The high validation accuracy and consistent performance across various metrics indicate that *FlameViT* is a robust model for wildfire smoke detection in satellite imagery.

4.8. Discussion

The *FlameViT* model significantly outperforms traditional CNN-based approaches, which achieved an accuracy of 86%. This improvement can be attributed to the inherent advantages of Vision Transformers, which are particularly well-suited for the task of wildfire smoke detection.

Vision Transformers excel in capturing long-range dependencies and intricate spatial relationships in high-resolution images. This capability is crucial for detecting wildfire smoke, which can be dispersed across large areas in satellite imagery. The self-attention mechanism in transformers allows the model to focus on different parts of the image, enabling it to identify subtle patterns of smoke that might be missed by CNNs.

A key aspect of *FlameViT* is its ability to use smoke as a proxy for wildfire detection. Detecting smoke is often a more reliable and early indicator of wildfires compared to detecting the fire itself, especially in satellite imagery where the fire may be obscured by vegetation or other elements. The model leverages the distinct visual features of smoke, such as its texture, spread, and translucency, which are effectively captured by the patch-based processing and attention mechanisms of the transformer architecture.

During the learning process, *FlameViT* learns to differentiate between smoke and other visually similar phenomena, such as fog, clouds, or mist. This differentiation is achieved through extensive training on a diverse dataset that includes various environmental conditions. The multi-head self-attention mechanism allows the model to focus on the context and finer details within each patch of the image, learning the unique characteristics of smoke in the presence of other elements. For instance, smoke from wildfires typically has a more dynamic and irregular pattern compared to the more uniform and stationary appearance of fog. The positional encoding also helps the model maintain spatial relationships, further aiding in distinguishing smoke patterns from other artifacts.

 doi:10.20944/preprints202408.1363.v2

The success of *FlameViT* in wildfire smoke detection has significant implications for fire detection and mitigation. Early detection of wildfire smoke allows for quicker response times, enabling firefighting teams to contain and extinguish fires before they spread uncontrollably. This can potentially save lives, protect property, and reduce economic losses.

Furthermore, the high accuracy and reliability of *FlameViT* make it a valuable tool for automated fire-fighting procedures. The model can be integrated into satellite monitoring systems, providing real-time alerts and detailed information about potential wildfire outbreaks. This can enhance the efficiency of firefighting efforts and improve overall disaster management strategies.

The relevance of these results extends beyond wildfire detection. The demonstrated effectiveness of Vision Transformers for high-resolution image analysis suggests that similar models could be applied to other remote sensing tasks, such as deforestation monitoring, agricultural assessment, and environmental protection.

5. Conclusions

In this study, we developed and evaluated *FlameViT*, a Vision Transformer-based model for detecting wildfire smoke in satellite images. The model demonstrated superior performance compared to traditional CNN methods, achieving a validation accuracy of 95%. This represents a significant improvement over the benchmark CNN model, which had an accuracy of 86%.

The *FlameViT* model's success can be attributed to its ability to capture long-range dependencies and intricate spatial relationships in high-resolution images, thanks to the self-attention mechanisms inherent in the transformer architecture. This capability is particularly important for detecting dispersed patterns of wildfire smoke, which are often challenging to identify with CNNs.

The implications of this research are profound. Early detection of wildfire smoke is crucial for timely and effective firefighting responses. By providing accurate and reliable predictions, *FlameViT* can help mitigate the devastating effects of wildfires, which have become increasingly severe due to climate change.

The integration of *FlameViT* into existing satellite monitoring systems could revolutionize wildfire detection and response strategies. Automated detection systems powered by *FlameViT* can provide real-time alerts, enabling rapid deployment of firefighting resources to contain fires before they spread. This can significantly reduce the loss of life, property damage, and economic impact caused by wildfires.

Beyond wildfire detection, the principles demonstrated in this research can be applied to a wide range of remote sensing tasks. The ability of Vision Transformers to handle high-resolution images with complex spatial dependencies makes them suitable for applications such as deforestation monitoring, crop health assessment, and environmental protection.

The success of *FlameViT* highlights the potential of Vision Transformers to advance the field of remote sensing and environmental monitoring. Future research can build on these findings by exploring the integration of additional data sources, such as weather data and topographical information, to further enhance the model's predictive capabilities.

Future work in the field of wildfire detection can explore several promising directions to further enhance the capabilities and applicability of models like *FlameViT*. One area of focus could be the integration of multimodal data sources, such as weather conditions, topographical maps, and historical fire data, to provide a more comprehensive context for wildfire prediction. By combining visual information from satellite images with these additional data streams, models could achieve higher accuracy and reliability, particularly in complex and diverse environmental conditions. Moreover, advancements in transfer learning and domain adaptation could enable the fine-tuning of pre-trained models on specific regions or seasons, improving their performance in varying geographic and temporal contexts.

Another significant area for future research is the development of real-time, deployable systems for automated wildfire detection and monitoring. This involves creating lightweight, efficient models that can run on edge devices or integrate seamlessly with satellite communication systems. Such systems

 doi:10.20944/preprints202408.1363.v2

could provide continuous surveillance and instant alerts, dramatically reducing the response time to emerging wildfires. Additionally, the use of explainable AI techniques could help in understanding and interpreting model predictions, making it easier for human operators and decision-makers to trust and act upon the system's outputs. Overall, the combination of enhanced data integration, real-time processing, and explainable AI promises to make wildfire detection systems not only more accurate but also more actionable and trustworthy in the fight against this growing environmental threat.

Furthermore, ongoing advancements in transformer architectures and training techniques can be leveraged to improve the performance and efficiency of models like *FlameViT*. As the field of machine learning continues to evolve, we can expect even more powerful tools for addressing the complex challenges posed by wildfires and other environmental threats.

In conclusion, *FlameViT* represents a significant advancement in the field of wildfire detection. By harnessing the power of Vision Transformers, we have developed a model that not only outperforms traditional methods but also offers a scalable and reliable solution for real-time wildfire monitoring. This research paves the way for more effective fire-fighting strategies and contributes to the broader goal of protecting our environment from the increasing threat of wildfires.

References

1. Bowman, D.M.; Balch, J.K.; Artaxo, P.; Bond, W.J.; Cochrane, M.A.; D'Antonio, C.M.; DeFries, R.S.; Johnston, F.H.; Keeley, J.E.; Krawchuk, M.A.; et al. Fire in the Earth system. *Science* **2009**, *324*, 481–484.
2. Mora, C.; McKenzie, T.; Gaw, I.L.; Dean, J.M.; von Hammerstein, H.; Knudson, T.A.; Setter, R.O.; Smith, C.Z.; Webster, K.M.; Patz, J.A.; et al. Broad threat to humanity from cumulative climate hazards intensified by greenhouse gas emissions. *Nature Climate Change* **2018**, *8*, 1062–1071.
3. Veraverbeke, S.; Rogers, B.M.; Goulden, M.L.; Jandt, R.R.; Miller, C.E.; Wiggins, E.B.; Randerson, J.T. Direct and indirect climate effects on spatial patterns of wildfires in boreal forest ecosystems. *Science advances* **2020**, *6*, eaay1121.
4. Jain, P.; Jain, P.K.; Chauhan, P. Review of forest fire detection techniques using wireless sensor network. *Materials Today: Proceedings* **2020**.
5. Xu, X.; Guo, Q.; Su, Y. GIS-based wildfire risk mapping and modeling for Mediterranean forests using logistic regression and multi-criteria decision analysis. *Forest Ecology and Management* **2016**, *368*, 163–172.
6. Radke, J.D.; Radke, J. Application of GIS technology in forest fire prevention and management. *International Journal of Geo-Information* **2019**, *8*, 394.
7. Yuan, Y.; Fang, H.; Deng, Z.; Li, S. Fire detection in UAV images using deep learning approach. *IEEE/ASME Transactions on Mechatronics* **2015**, *20*, 2893–2904.
8. Zhao, H.; Wang, Z.; Xu, B.; Liu, Q.; Zhang, Y. UAV-based remote sensing for forest fire monitoring and assessment. *Journal of Remote Sensing* **2018**, *22*, 578–589.
9. Hartung, C.; Han, R.; Seielstad, C. Fire risk assessment using wireless sensor networks. *2006 Fourth Annual IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)* **2006**, pp. 13–17.
10. Mao, X.; Xie, Q.; Tang, S. Wireless sensor networks for fire risk prediction and forest fire detection. *Sensors* **2019**, *19*, 1441.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
13. Liu, Y.; Lapata, M. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345* **2019**.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
15. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877* **2020**.
16. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* **2016**.
17. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* **2019**, *32*, 5753–5763.

18. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. *arXiv preprint arXiv:1810.04805* **2018**.
19. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* **2019**.
20. Dong, L.; Xu, S.; Xu, B. Speech-Transformer: A no-recurrence sequence-to-sequence model for speech recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* **2018**, pp. 5884–5888.
21. Li, S.; Jin, X.; Xie, C.; Jiang, H.; Pan, H. Enhancing time-series momentum strategies using deep neural networks: A systematic learning framework. *Journal of Financial Data Science* **2019**, *1*, 45–64.
22. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.
23. Rao, R.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; et al. MSA Transformer. *bioRxiv* **2021**.
24. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. *arXiv preprint arXiv:2012.09841* **2021**.
25. Fan, L.; Xie, D.; Zeng, W.; Wang, H.; Pu, S. Learning to drive through deep reinforcement learning. *IEEE Transactions on Vehicular Technology* **2021**, *70*, 1062–1073.
26. Chen, L.; Lu, K.; Rajeswaran, A.; Lee, K.; Grover, A.; Laskin, M.; Srinivas, A.; Abbeel, P.; Mordatch, I.; et al. Decision transformer: Reinforcement learning via sequence modeling. *arXiv preprint arXiv:2106.01345* **2021**.
27. Li, L.H.; Su, W.; Xiong, C.; et al. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* **2019**.
28. Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; Schmid, C. VideoBERT: A joint model for video and language representation learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision* **2019**, pp. 7464–7473.
29. Child, R.; Gray, S.; Radford, A.; Sutskever, I. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509* **2019**.
30. Kitaev, N.; Kaiser, Ł.; Levskaya, A. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451* **2020**.
31. Choromanski, K.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Belanger, D.; Colwell, L.; Weller, A. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794* **2020**.
32. Wang, S.; Li, B.Z.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768* **2020**.
33. Pettorelli, N.; Laurance, W.F.; O'Brien, T.G.; Wegmann, M.; Nagendra, H.; Turner, W. Satellite remote sensing for applied ecologists: opportunities and challenges. *Journal of Applied Ecology* **2013**, *50*, 830–841.
34. Liu, Z.; Yang, J.; Chang, Y.; Jiao, L. Assessment of forest fire risk based on fuzzy AHP and fuzzy comprehensive evaluation. *Ecological Modelling* **2015**, *297*, 42–50.
35. Xi, W.; Li, J. Integrating multi-source data to improve forest fire detection based on random forests. *Remote Sensing* **2019**, *11*, 297.
36. Srivastava, P.K.; Han, D.; Rico-Ramirez, M.A.; Bray, M.; Islam, T.; Dai, Q. Deep learning for precipitation prediction: Towards better accuracy. *Meteorological Applications* **2020**, *27*, e1874.
37. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, *25*, 1097–1105.
38. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* **2016**, pp. 770–778.
40. Dutta, R.; Sgurr, G. Improving fire detection using an ensemble of convolutional neural networks trained with satellite-based image patches. *Remote Sensing Letters* **2016**, *7*, 1215–1224.
41. Mendonça, E.A.d.S.; Martinez, A.L. Wildfire detection in the brazilian amazon rainforest using convolutional neural networks and satellite images. *Remote Sensing* **2020**, *12*, 3172.
42. Farasin, I.; Anedda, M.; Fanni, A.; Martis, L. Deep learning techniques for wildland fires analysis through aerial images. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* **2017**, pp. 1068–1071.
43. Raj, S.; Suriyalakshmi, K.; Srinivasan, K. Detection of wildfires using recurrent neural networks. *International Journal of Disaster Risk Reduction* **2020**, *49*, 101745.
44. Zhang, J.; Zhang, Z.; Ma, J.; Wang, L. Enhancing spatial attention using dual attention mechanism for wildfire detection in satellite images. *Remote Sensing Letters* **2019**, *10*, 903–912.

 doi:10.20944/preprints202408.1363.v2

45. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv preprint arXiv:2102.04306* **2021**.
46. Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; Jégou, H. Training data-efficient image transformers and distillation through attention. *arXiv preprint arXiv:2012.12877* **2021**.
47. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.
48. Xiang, W.N.; Clarke, K.C. Spatial modeling of wildland fire risk in the wildland-urban interface using support vector machines and geographic information systems. *Environmental Modelling & Software* **2016**, *83*, 207–219.
49. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Rabinovich, A. Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition* **2015**, pp. 1–9.
50. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* **2017**, pp. 4700–4708.