Check for updates

JGR | Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2024JH000409

Comparative and Interpretative Analysis of CNN and Transformer Models in Predicting Wildfire Spread Using Remote Sensing Data

Yihang Zhou[1], Ruige Kong[2], Zhengsen Xu[3], Linlin Xu[4], and Sibo Cheng[5]

[1]Department of Earth Science and Engineering, Imperial College London, London, UK, [2]Department of Engineering, University of Cambridge, Cambridge, UK, [3]Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON, Canada, [4]Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada, [5]CEREA, ENPC, EDF R&D, Institut Polytechnique de Paris, Palaiseau, France

Key Points:

- Our comprehensive analysis compared four prevalent deep learning models for predicting wildfires using a decade-long data set from the U.S
- Our comparative analysis demonstrates that UNet and Transformer-based Swin-UNet outperform conventional CNN methods in wildfire prediction
- We enhanced model transparency and understood their characteristics, providing insights for optimal model selection and future improvements

Correspondence to:

S. Cheng,
sibo.cheng@enpc.fr

Citation:

Zhou, Y., Kong, R., Xu, Z., Xu, L., & Cheng, S. (2025). Comparative and interpretative analysis of CNN and transformer models in predicting wildfire spread using remote sensing data. *Journal of Geophysical Research: Machine Learning and Computation*, *2*, e2024JH000409. https://doi.org/10.1029/2024JH000409

Received 20 SEP 2024
Accepted 27 FEB 2025

Abstract Facing the escalating threat of global wildfires, numerous computer vision techniques using remote sensing data have been applied in this area. However, the selection of deep learning methods for wildfire prediction remains uncertain due to the lack of comparative analysis in a quantitative and explainable manner, crucial for improving prevention measures and refining models. This study aims to thoroughly compare the performance, efficiency, and explainability of four prevalent deep learning architectures: Autoencoder, ResNet, UNet, and Transformer-based Swin-UNet. Employing a real-world data set that includes nearly a decade of remote sensing data from California, U.S., these models predict the spread of wildfires for the following day. Through detailed quantitative comparison analysis, we discovered that Transformer-based Swin-UNet and UNet generally outperform Autoencoder and ResNet, particularly due to the advanced attention mechanisms in Transformer-based Swin-UNet and the efficient use of skip connections in both UNet and Transformer-based Swin-UNet, which contribute to superior predictive accuracy and model interpretability. Then we applied XAI techniques on all four models, this not only enhances the clarity and trustworthiness of models but also promotes focused improvements in wildfire prediction capabilities. The XAI analysis reveals that UNet and Transformer-based Swin-UNet are able to focus on critical features such as "Previous Fire Mask", "Drought", and "Vegetation" more effectively than the other two models, while also maintaining balanced attention to the remaining features, leading to their superior performance. The insights from our thorough comparative analysis offer substantial implications for future model design and also provide guidance for model selection in different scenarios. The source code for this project is publicly available as open source on Zenodo (Y. Zhou et al., 2024, https://doi.org/10.5281/zenodo.14286931).

Plain Language Summary As wildfires increase globally, predicting their occurrence accurately and understandably has become more critical than ever. This study compared advanced computer models using a decade of space-based data to predict next-day wildfire risks in the United States. We focused on two types of models: Convolutional Neural Networks and Transformer models. A key aspect of our research was not only determining which model predicts wildfires more accurately but also understanding how these models make their predictions. This is where XAI plays a crucial role. XAI helps us explore these complex models to see how they process and interpret data, ensuring that the predictions they make are both reliable and transparent. Through detailed comparisons and analyses, our findings highlight significant differences in model performance and their approaches to interpreting data. Emphasizing both accuracy and explainability, this study enhances our ability to select and refine the best models for predicting wildfires, offering crucial insights that could improve prevention strategies and advance wildfire prediction technology.

© 2025 The Author(s). *Journal of Geophysical Research: Machine Learning and Computation* published by Wiley Periodicals LLC on behalf of American Geophysical Union.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

1. Introduction

1.1. Background

Wildfires pose a significant threat to ecosystems, property, and human life worldwide (Bousfield et al., 2023; Nolan et al., 2022). Accurate and timely prediction of these natural disasters is essential for effective prevention and management strategies. In recent years, the emergence of remote sensing technology has revolutionized wildfire monitoring and prediction. Remote sensing data, including a range of spectral, spatial, and temporal

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

resolutions, provide critical information on vegetation status, moisture content, topography, and other environmental factors instrumental in wildfire development and spread.

The integration of deep learning models with remote sensing data has emerged as a promising approach to wildfire prediction and analysis. Deep learning models, such as autoencoders (Huot et al., 2022b), ResNet (He et al., 2016), UNet (Ronneberger et al., 2015), and Vision Transformers (ViT) (Dosovitskiy et al., 2020), have demonstrated substantial capabilities in processing complex spatial and temporal data, offering significant advantages in efficiency and accuracy over traditional statistical methods. These models can identify subtle patterns and correlations in large data sets, facilitating more accurate predictions of wildfire occurrences (Khryashchev & Larionov, 2020; Suwansrikham & Singkhamfu, 2023) and behavior (Ivek & Vlah, 2022). For example, Al-Dabbagh and Ilyas (2023) processed uni-temporal Sentinel-2 imagery using UNet with ResNet50, achieving an F1-score of 98.78% and an IoU of 97.38% in the semantic segmentation of wildfire-affected areas. The data set used in this study, captured by the Multi Spectral Instrument sensor, has a spatial resolution of 10 m and consists of data from five non-continuous days in September 2021, highlighting its temporal and spatial complexity. Similarly, Suwansrikham and Singkhamfu (2023) utilized the ViT model to process UAV aerial photography data, achieving the highest accuracy of 88.03% in forest fire detection. This data set consists of 56 historical fire georeferenced perimeters from the period of 2014–2016, with a spatial resolution of 30 m, demonstrating the capability of the.

Transformer-based model to handle and interpret complex spatial data effectively.

However, the complexity of deep learning models often leads to a "black box" problem where the decision-making process is neither transparent nor understandable (Castelvecchi, 2016). This lack of explainability can be a significant barrier, especially in high-stakes scenarios such as wildfire prediction, where understanding the rationale behind predictions is crucial for trustworthy and actionable insights and for guiding subsequent modeling and decision-making (Girtsou et al., 2021; Kondylatos et al., 2022). For example, Abdollahi and Pradhan (2023) underscored the need for explainability in wildfire prediction models as their outputs guide natural resource management plans, necessitating an understanding of the underlying logic by decision-makers. Fan et al. (2024) also emphasized the importance of explainability when facilitating more effective and informed forest fire management strategies. Therefore, in this paper, to address the critical role of model explainability, we employ a suite of interpretability methods: SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017), Gradient-weighted Class Activation Mapping (Grad-CAM) (Sundararajan et al., 2017), and Integrated Gradients (IG) (Selvaraju et al., 2017), to dissect and illuminate the decision-making process of our predictive models. Beyond merely applying these tools, we conduct a detailed analysis to uncover the most influential features for wildfire propagation, and how these models prioritize the features in the data set, helping us explore their characteristics and reveal their strengths and weaknesses.

Through SHAP, we quantify the impact of each environmental factors on model predictions across the entire data set, laying a foundation for transparent model evaluation. Grad-CAM complements this by providing visual explanations that highlight critical areas within input images, thus allowing visual validation of the model's focus. Integrated Gradients extends this analysis by attributing the contributions of specific features to the prediction outcomes for individual samples in the data set. Collectively, these methods enable us to pinpoint critical environmental factors, clarify their relative importance and interactions within the models. This approach not only augments the transparency and reliability of the four models we implemented but also fosters targeted enhancements in wildfire prediction capabilities. The overall workflow of this study is illustrated in Figure 1.

The primary objective of this research is to compare the efficiency of various deep learning models, specifically CNNs and Transformer-based model, in predicting wildfires using remote sensing data. These two categories of models possess inherently different characteristics, that is, CNNs, such as autoencoders, ResNet, and UNet are good at capturing spatial hierarchies in imagery, while Transformers are efficient at modeling long-range dependencies and integrating context more comprehensively. We hypothesize that, these architectural differences will lead to varying performance in wildfire prediction tasks. Furthermore, by integrating explainability methods, such as SHAP values and Grad-CAM, we aim to offer valuable insights into the models' decision-making processes. Specifically, how they weigh and interpret different features in remote sensing data to arrive at their predictions. We expect this exploration to deepen our understanding of the models' predictive capabilities and the rationale behind their forecasts, thereby enhancing both the transparency and reliability of wildfire forecasting models.

10.1029/2024JH000409

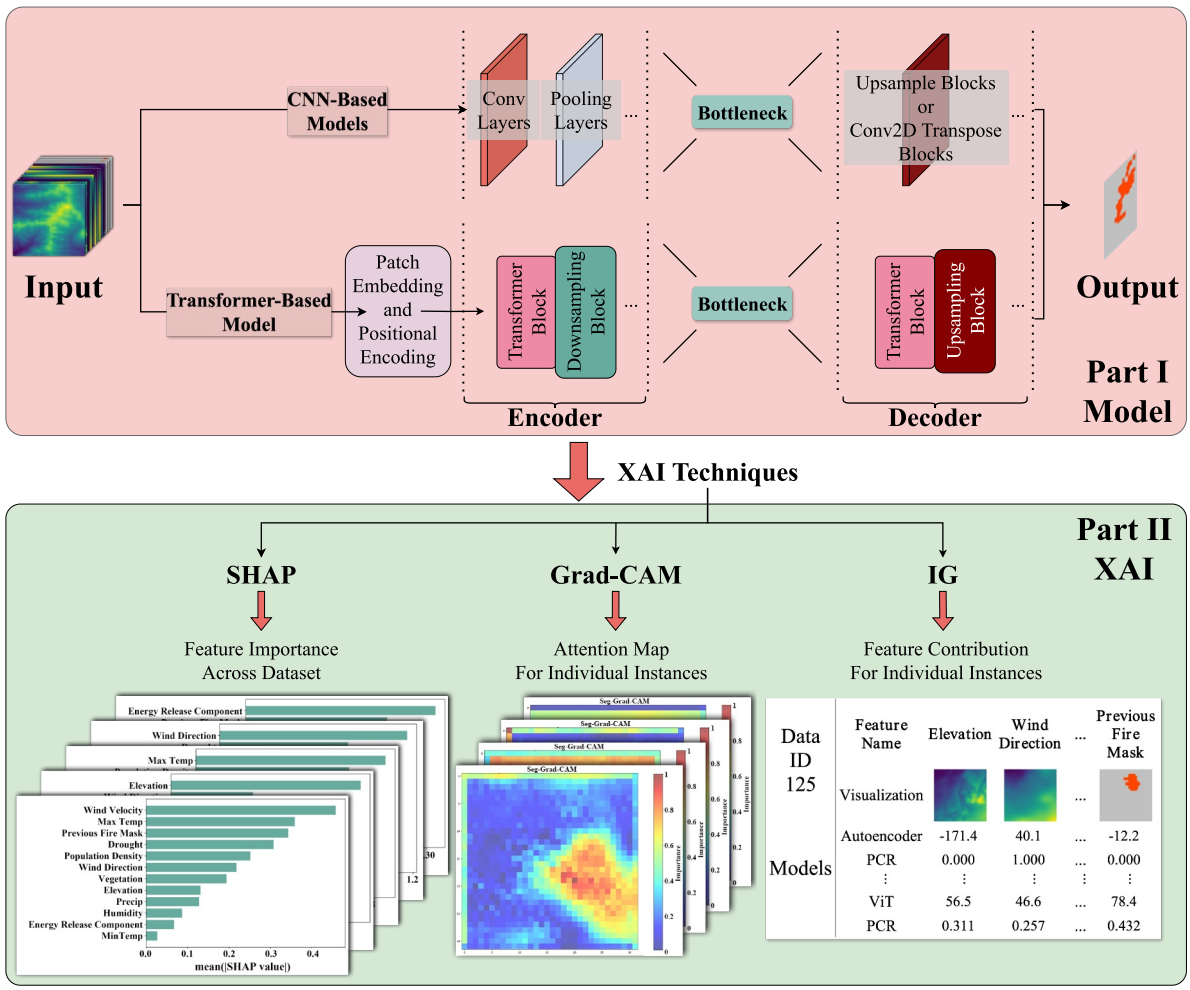


Figure 1. Workflow of this analysis. Starting with the input data (including meteorological data, multispectral data, terrain data, and so on), it is processed by two types of models: CNN-based models and a Transformer-based model. For all models, we further applied three explainability analysis techniques: SHAP, Grad-CAM, and Integrated Gradients (IG), to explore and interpret the decision-making processes of the models. In the table of IG, "Data ID 125" refers to the 125th sample in the data set. Positive contribution ratio stands for Positive Contribution Rate, a metric indicating the contribution of a specific feature.

1.2. Related Works

The evolution of wildfire prediction methods has significantly transitioned from reliance on traditional statistical methods to the incorporation of complex machine learning algorithms (Jain et al., 2020; Pham et al., 2022). Initially, predictions were primarily based on empirical models utilizing meteorological data and historical fire records, which despite their utility, often faced limitations such as time inefficiency and a lack of flexibility for parameter tuning (Perry, 1998). In addition to these empirical models, physical-based models have also been widely employed in wildfire prediction, such as The Wildland Fire Dynamics Simulator (WFDS) (Mell et al., 2007), UoC-R (X. Zhou et al., 2005), and UoS (Asensio & Ferragut, 2002). These models, which are grounded in the fundamental chemistry and physics of combustion and fire spread, aim to simulate the behavior and dynamics of wildland fires more realistically (Sullivan, 2009). The emergence of remote sensing technologies, such as satellite imagery and aerial photography, marked a considerable advancement by enriching the data set available for analysis and improving the timeliness and reliability of wildfire predictions (Campbell & Hossain, 2022; Rashkovetsky et al., 2021).

Deep learning models, including CNNs like autoencoders, ResNet, UNet, and Transformer-based models, offer unique strengths in analyzing complex spatial and temporal patterns. Compared to traditional machine learning methods, deep learning approaches have been shown to achieve higher performance in wildfire prediction. For

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

instance, in the comparison by Huot et al. (2022b), Huot et al. (2022b) autoencoders significantly outperformed traditional machine learning methods, with notable improvements in several metrics, demonstrating the ability of deep learning to capture more complex patterns and improve prediction accuracy. Autoencoders are good at reducing dimensionality, denoising data, and discovering hidden features, making them effective for handling complex spatial and temporal data (Cheng et al., 2022; Zhai et al., 2018). ResNet's deep learning framework excels in learning detailed features through its extensive layers (He et al., 2016), while UNet combines context and localization, making it ideal for spatial tasks, such as the mapping of wildfire spread (Hodges & Lattimer, 2019). In contrast, Transformer-based model uses self-attention mechanisms, focusing on global data relationships rather than spatial proximity (Pan et al., 2023), which is effective for understanding widespread environmental influences on wildfires. Additionally, Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) are commonly used for wildfire prediction due to their ability to handle sequential data and capture temporal dependencies. However, these models were not included in our comparative analysis because our focus is on evaluating the effectiveness of CNNs and Transformer models in handling the spatial complexities present in remote sensing data for wildfire prediction.

Recent studies Khanmohammadi et al. (2022), Zhong et al. (2023), and Cheng et al. (2023)) have advanced the application of deep learning in wildfire prediction significantly, as summarized in Xu et al. (2024). For instance, deep learning models such as CNNs, RNNs and Transformers have been employed using satellite data to predict wildfire severity and danger, as reviewed by Guo et al. (2023). Furthermore, Ji et al. (2024) proposed a model combining ConvLSTM networks with spatial feature extraction from U-Net and SKNet (X. Li et al., 2019) networks, effectively enhancing global wildfire danger predictions. Other notable efforts include the work by Laube and Hamilton (2021), who collected SaskFire data set and utilized ResNet, achieving a precision of 0.25 and a recall of 0.80 on a highly imbalanced data set. Additionally, Kondylatos et al. (2022) explored CNNs and Transformer-based architectures, underscoring their effectiveness in forecasting wildfire danger. These differences highlight the necessity for comparative analysis to identify the most effective models for wildfire prediction.

1.3. Contribution

To address the “black box” problem more comprehensively and achieve a deeper understanding of the decision-making process, XAI techniques such as SHAP, Grad-CAM, and IG are introduced. In the field of geoscience, researchers have acknowledged the significance of interpretability and have applied it in various domains, such as climate prediction (Bommer et al., 2023; Mamalakis et al., 2022) and drought forecasting (Dikshit & Pradhan, 2021). Researchers have also attempted to incorporate explainability into wildfire scenarios. For instance, Ahmad et al. (2023) designed FireXnet, achieving a 98.42% test accuracy, which outperforming models like VGG16 and DenseNet201, and also integrated SHAP for explainability. Similarly, Qayyum et al. (2024) employed a transformer encoder-based approach for wildfire prediction, using SHAP analysis to elucidate the connection between predictor variables and model performance. Overall, these studies highlight the growing precision and explanatory capabilities of deep learning models in wildfire prediction. However, the effectiveness and applicability of these methods in enhancing model transparency, especially within the context of remote sensing for wildfire prediction, have not been fully explored. For example, they often restrict explainability to general dataset-level analysis without further detailed examination, such as analyzing the outputs of each layer of the models and on each sample in the data set, which could potentially offer deeper insights into the model's behavior and improve our understanding of its decision-making processes.

To address the identified research gaps and challenges in our study, we have undertaken a comprehensive approach that is outlined in several key areas. Firstly, we conducted a thorough quantitative analysis to compare the efficiency and performance of CNNs and Transformer models in predicting wildfires. This evaluation includes not just performance metrics but also efficiency aspects such as the number of parameters and GFLOPs, revealing distinct strengths and weaknesses of each model type. This comprehensive comparison serves as a foundation for understanding the trade-offs between model complexity and predictive accuracy. Secondly, we integrated advanced interpretability methods to investigate the decision-making processes behind the models' predictions. Our dual approach of analyzing the entire data set through SHAP and assessing individual samples via Grad-CAM and IG clarifies how models prioritize and balance different environmental factors. This detailed analysis not only enhances model interpretability and transparency but also improves the credibility of wildfire prevention measures based on model predictions. Lastly, we balanced model performance and interpretability to

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

provide valuable guidance for selecting the most suitable model for wildfire prediction, taking into account varying needs such as accuracy, recall, real-time performance, and computational resource demands. The insights obtained from our study also provide a clear direction for future enhancements in wildfire prediction models, by identifying critical factors that influence model accuracy and interpretability.

In summary:

1. We conducted a comprehensive analysis comparing models, including CNNs(a baseline Autoencoder proposed by Huot et al. (2022b), ResNet, and UNet) and a Transformer-based model, all implemented from scratch, on their effectiveness and efficiency in wildfire prediction. Moreover, models such as UNet and Transformer-based Swin-UNet demonstrated superior performance compared to the baseline Autoencoder.
2. We enhanced model transparency and interpretability through an integrative XAI approach incorporating SHAP, Grad-CAM, and IG. This contributes to advancing the application of XAI in predicting wildfires.
3. We provided guidance for selecting suitable wildfire prediction models and outlined key areas for future research and enhancements.

In Section 2, we detail the data set utilized in this study, describe the architectures of the deep learning models employed, and outline the theoretical framework of the XAI methods applied. Section 3 presents the evaluation metrics and offers a performance analysis based on the corresponding experimental results. Section 4 provides an in-depth interpretability analysis, examining each model feature by feature. Finally, the paper concludes with a summary and future directions in Section 5.

2. Methodology

2.1. Data Set Description

Considering that both CNNs and Transformer-based models excel at processing complex spatial and temporal data, it is critical that our data set is rich, multifaceted, and high-precision to fully leverage these characteristics of the models. Using primarily the "Next Day Wildfire Spread" data set (Huot et al., 2022b), aggregated via Google Earth Engine (Gorelick et al., 2017), this study harnesses a comprehensive, multivariate collection of historical wildfire data across the United States. This data set is unparalleled, integrating a decade's worth of satellite observations, standardized to a 1 km resolution, including previous and current fire masks from key sensors such as the Visible Infrared Imaging Radiometer Suite (VIIRS) and the Shuttle Radar Topography Mission (SRTM) (Didan & Barreto, 2018; Farr et al., 2007).

The data set includes 12 essential input features that capture environmental, meteorological, and anthropogenic factors influencing wildfire behavior. These features are derived from various sources with different native spatial resolutions and temporal characteristics, all standardized to a 1 km resolution for consistency. Specifically, *Elevation* is derived from SRTM data (Farr et al., 2007) at a native resolution of 30 m, providing critical terrain information. *Wind direction* and *Wind velocity* are obtained from the GRIDMET data set (Tavakkoli Piralilou et al., 2022) at 4 km resolution, representing daily atmospheric conditions that affect fire spread and intensity. *Minimum temperature*, *Maximum temperature*, *Humidity*, and *Precipitation* are also sourced from GRIDMET.

(Tavakkoli Piralilou et al., 2022) at 4 km resolution, capturing daily weather conditions influencing fuel moisture and fire ignition likelihood. The *Drought* variable is derived from the U.S. Drought Monitor data, sampled every 5 days at 4 km resolution, integrating indicators such as precipitation, soil moisture, and streamflow to classify drought severity, which is crucial for assessing long-term dryness conditions that elevate wildfire risk. *Vegetation* information is obtained from the Normalized Difference Vegetation Index (NDVI) provided by VIIRS (Didan & Barreto, 2018) at a native resolution of 0.5 km, sampled every 8 days, indicating fuel availability and condition vital for predicting fire spread. *Population Density* is sourced from the Gridded Population of the World (GPWv4) data set, updated every 5 years at 1 km resolution, serving as a proxy for human activity that may influence ignition rates and fire management practices. The *Energy Release Component* (ERC), from the National Fire Danger Rating System, is available daily at 1 km resolution, representing potential energy release per unit area in the flaming front of a fire, reflecting fuel dryness and potential fire intensity. Lastly, the *Previous fire mask* represents fire locations at time t, obtained from MOD14A1 fire mask composites (Giglio & Justice, 2015) at 1 km resolution, providing historical fire occurrence information essential for understanding fire progression.

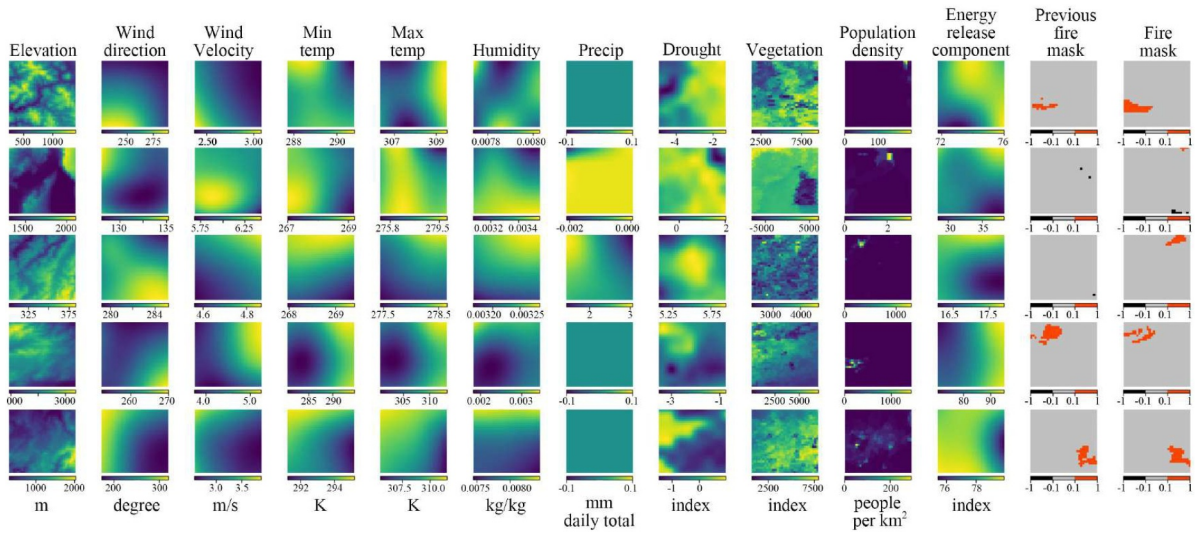


Figure 2. Visualized data set (Huot et al., 2022b). Each row represents a sample from the data set, displaying all input features and the corresponding output. The first 12 columns represent the input features, such as "Elevation", "Wind direction", and so on. The last column represents the label, where red indicates the presence of fire, gray signifies no fire, and black is used for uncertain labels, such as instances obscured by cloud coverage or other unprocessed data.

Each of these features is standardized to a spatial resolution of 1 km to ensure consistency across different data sources and facilitate effective integration into our models. The data set spans from 2012 to 2020, utilizing daily snapshots or the most recent data available before the fire event (t), thereby ensuring that our models are informed by the conditions most likely to influence wildfire behavior. Most variables represent conditions immediately prior to the fire event, not long-term means. For example, meteorological variables and ERC are daily values capturing immediate conditions, while drought and vegetation indices use the most recent data before t to approximate current conditions. Population density and elevation are treated as static over the study period.

This selection offers an opportunity to investigate wildfire spread with a level of detail and temporal resolution that is unmatched by other publicly available resources. Unlike existing fire data sets that primarily focus on burn areas without offering comprehensive environmental context or the necessary 2-D, day-by-day progression for accurate fire spread analysis (such as FRY (Laurent et al., 2018), Fire Atlas (Andela et al., 2019), MOD14A1 V6 (Giglio & Justice, 2015), GlobFire (Artés et al., 2019), VNP13A1 (Didan & Barreto, 2018), and GRIDMET (Tavakkoli Piralilou et al., 2022)), our chosen data set fills this gap by incorporating a wide array of variables critical for advanced predictive modeling. The output is the next day's fire mask, representing fire occurrences at time $t + 1$.

The data set is constructed from TFRecord files and configured with a batch size of 100, accommodating 12 input channels and 1 output channel. Some data set visualizations are shown in Figure 2. Our study involves a thorough data preprocessing stage, essential for the effective training of our models. Each feature is normalized based on predetermined statistics: minimum, maximum, mean, and standard deviation. To prepare the data, we used a random-cropping method, which ensures uniform-sized inputs for the model by extracting relevant sections from the images.

2.2. Models

2.2.1. Autoencoder Architecture and Implementation

The autoencoder architecture, fundamental for dimensionality reduction in deep learning, was primarily inspired by Hinton and Salakhutdinov (2006). Our model was adapted from the baseline model detailed in Huot et al. (2022b), which has 12 input channels and 1 output channel. The architecture of our model incorporates a sequence of convolutional layers, designed with dimensions following the sequence $[64, 128, 256, 256, 256]$. To complement the convolutional layers, each is succeeded by a pooling layer, specifically of size 2×2, ensuring a structured reduction in spatial dimensions while retaining critical feature information. This configuration is

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

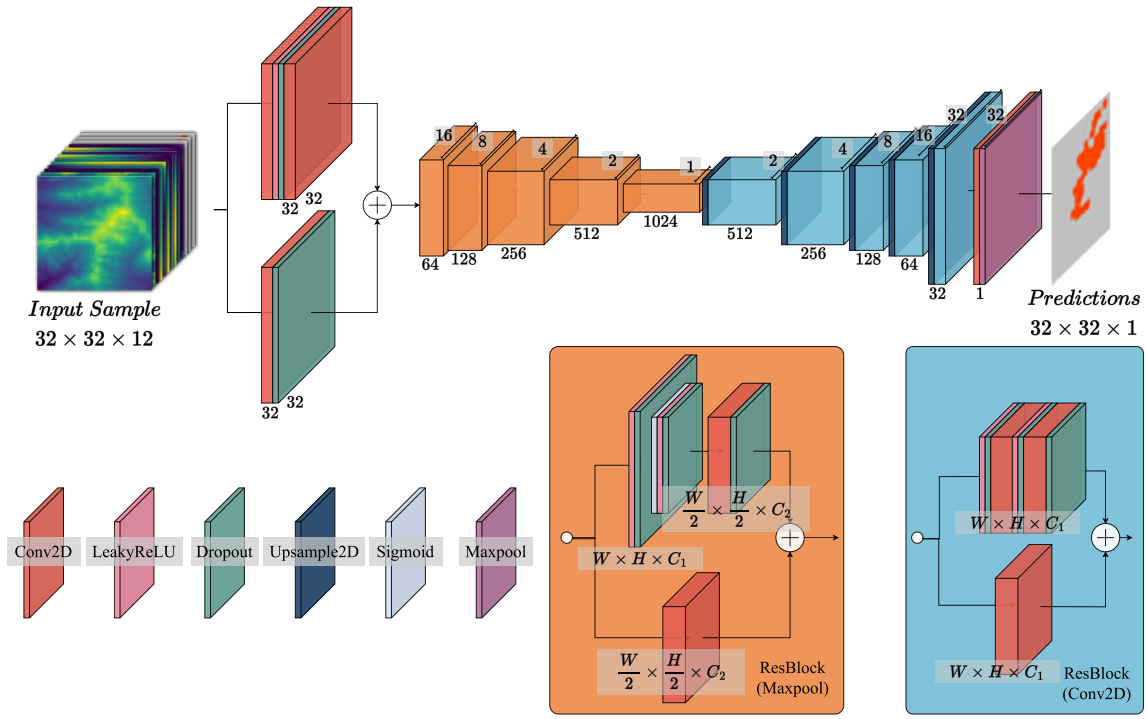


Figure 3. Detailed structure for baseline Autoencoder.

strategically optimized to enhance the model's ability to capture and process the hierarchical spatial features essential for accurate wildfire prediction. The detailed structure of the Autoencoder can be seen in Figure 3.

2.2.2. ResNet Architecture and Implementation

The ResNet (Residual Network) architecture, a significant advancement in deep learning, is renowned for its deep structure that can run hundreds of layers. It addresses the vanishing gradient problem through "skip connections", also known as "residual connections", which allow the gradient to flow through the network without attenuation (He et al., 2016). ResNet's success in image recognition tasks has established it as a benchmark model in the field. In our study, we adapted the ResNet architecture, specifically using the ResNet50 variant, to predict wildfires from remote sensing data. To enhance our model's ability to extract features crucial for predicting wildfires, we added an additional convolutional layer before the ResNet encoder. This adjustment aims to improve the model's initial processing of input data. Further modifications include adopting the decoder structure from the existing baseline autoencoder model, integrating it with the ResNet encoder to form an efficient encoder-decoder setup. This setup is specifically designed to handle spatial features effectively, which are vital for accurate wildfire prediction. The decoder is structured to invert the process of the encoder, using reversed layers and pooling operations to reconstruct the output from encoded data. A customized architecture of the ResNet can be seen in Figure 4. The integration of ResNet with a decoder from an autoencoder model in our implementation is a novel approach. This combination leverages ResNet's powerful feature extraction capabilities and the decoder's efficient spatial reconstruction, making it highly suitable for the complex task of wildfire prediction from varied and intricate remote sensing data.

2.2.3. UNet Architecture and Implementation

The UNet architecture is distinguished by its U-shaped structure, featuring a contracting path for context capture and an expansive path for precise localization (Ronneberger et al., 2015). In particular, the UNet is distinguished in its ability for detailed segmentation tasks using a limited data set. This allows us to accurately segment complex spatial patterns in satellite imagery and facilitates the identification of wildfire-prone areas. In our wildfire prediction study, we adapted the UNet model to process remote sensing data. Our model is initiated with an input layer designed to accommodate data dimensions of 32 × 32 across 12 channels, setting the foundation for

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10.1029/2024JH000409

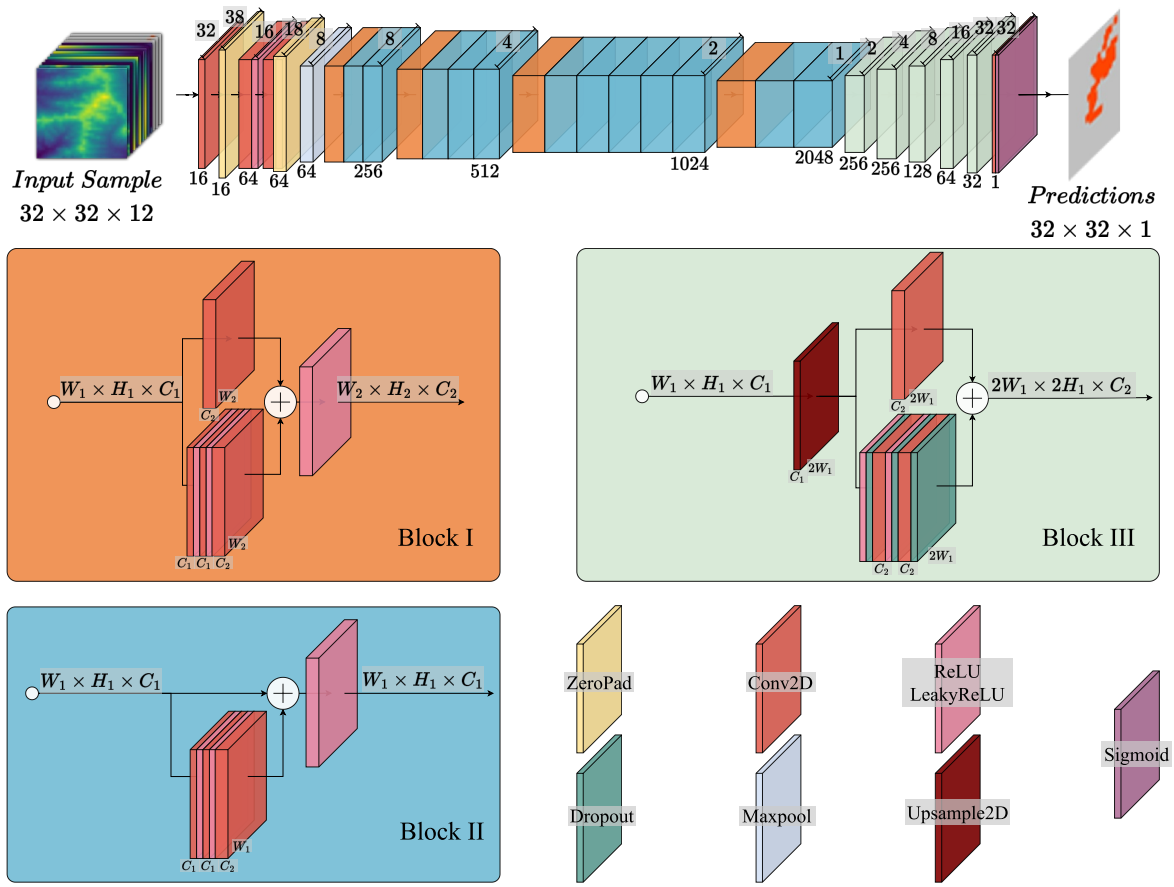


Figure 4. Detailed structure of ResNet, where skip connections are confined within individual blocks, which is different from the UNet that utilizes skip connections spanning from encoder to decoder.

complex feature processing. The architecture progresses through a series of Convolutional 2D (Conv2D) layers, where the filter sizes gradually increase from 32 to 512. This escalation allows for a hierarchical extraction of features, ensuring a detailed understanding of the input data. Following each Conv2D layer, batch normalization is applied to stabilize the training process, addressing internal covariate shift and speeding up convergence. Spatial dimension reduction is achieved through max pooling in the contracting path, while the expanding path uses transposed convolutions for detailed spatial feature mapping. To take into account the risk of overfitting, dropout layers are placed in the deeper sections of the model. In our model, feature maps from the contracting path are precisely merged with those in the expansive path. The merging process preserves vital spatial details throughout the network, ensuring both the context and localization accuracy are enhanced for reliable wildfire prediction. The detailed structure of UNet can be seen in Figure 5.

2.2.4. Transformer-Based Approaches Architecture and Implementation

Transformer-based approaches introduce a novel method for image analysis by dissecting images into patches and processing them similarly to words in a sentence (Dosovitskiy et al., 2020). This technique allows these models to understand and connect different parts of an image globally, unlike traditional models that focus on local areas. In the context of wildfire prediction, it enables the model to detect subtle yet significant patterns across vast landscapes, such as changes in vegetation dryness or unusual temperature variations, which are key indicators of potential wildfire outbreaks. By capturing these comprehensive spatial relationships, transformer-based models can predict wildfires with higher accuracy, contributing to more effective monitoring and management of wildfire risks. Swin-UNet, an innovative adaptation of this architecture, is customized for segmentation tasks, enhancing the traditional UNet structure with Swin Transformer blocks (Cao et al., 2021). This architecture is ideal for handling hierarchical features, a critical aspect in processing remote sensing data for wildfire prediction.

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

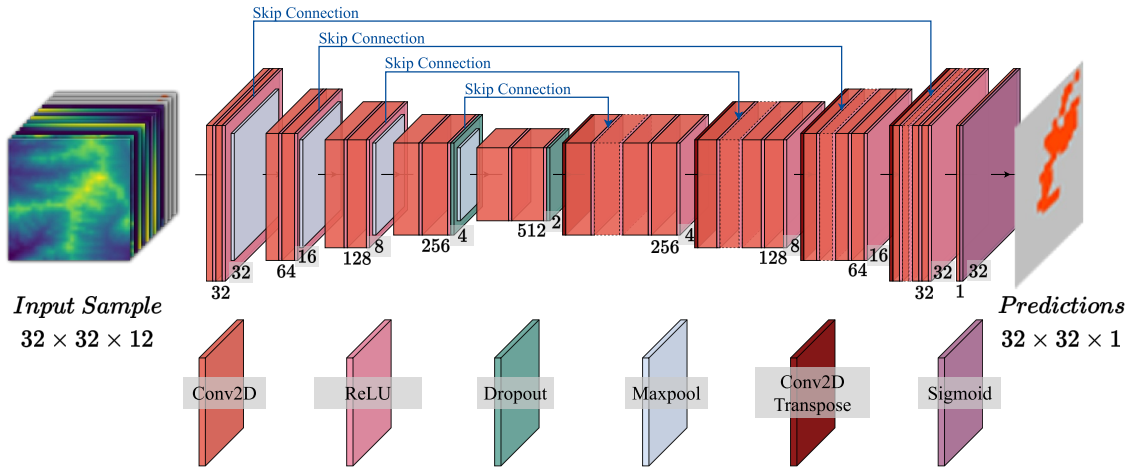


Figure 5. The detailed structure of UNet, featuring skip connections that extend from the encoder to the decoder, represented by blue lines in the illustration.

In our Swin-UNet setup for wildfire prediction, we use a Swin Transformer for both the encoder and decoder parts. The model begins with a 16-channel convolutional layer to align with the initial convolutional layers of three other models, ensuring consistent heatmaps generated by Grad-CAM. Following this, a downsampling block with 128 channels is employed to extract robust features. It has a depth of 4, including three down/upsampling levels and an extra bottom level for processing features at multiple scales. Each stage has two Swin Transformer blocks that use Window-based Multi-head Self-Attention (W-MSA) and Shifted Window-based Multi-head Self-Attention (SW-MSA) to capture both local and global details. Window-based Multi-head Self-Attention focuses on capturing spatial relationships within local windows, while SW-MSA shifts window positions to capture cross-window connections, enhancing global context integration. The 2×2 patch sizes help extract detailed image patches, which is important for capturing key details often missed by regular CNNs. The different attention heads and window sizes in the Swin Transformer blocks allow the model to analyze various spatial scales in an image, from large landscape features to small changes in vegetation or terrain. Layer Normalization (LN) is used to normalize the inputs across features for each layer, improving training stability and speed. The Multi-Layer Perceptron (MLP) introduces non-linearity through fully connected layers, helping the network learn complex features. During the upsampling process, the dense layers placed in conjunction with patch expanding layers increase the feature dimensions back to their original size. This makes Swin-UNet very good for remote sensing data, combining effective spatial resolution and accurate segmentation tasks. The detailed architecture is shown in Figure 6.

2.2.5. Training Setup

In the configuration of our model, both the input features and the target fire mask were established with spatial dimensions of 32×32. The optimization process was executed using the Adam optimizer, which was configured with a learning rate of $\alpha = 0.0001$ and first and second-moment exponential decay rates set to $\beta_1 = 0.9$, $\beta_2 = 0.999$ (Kingma & Ba, 2015).

To address the challenges posed by class imbalances and to focus more precisely on the classes most relevant for wildfire prediction, we adopted a custom *masked weighted cross-entropy* loss function, as outlined in Huot et al. (2022b). Furthermore, to mitigate the risk of overfitting, we diverged from the original approach by incorporating an early stopping mechanism that concludes training if no improvement is observed over 30 epochs, in contrast to the original paper's use of a fixed 1,000 epochs. While both studies employ the Adam optimizer, the precise optimization hyperparameters remain distinct and are customized in our study to address the challenges of wildfire prediction. The input features retain a resolution of 32×32 as established in the original framework, ensuring consistency in data representation. The Google Cloud computing T4 GPU is used for training, and the batch size is set to 100 for all the approaches.

The implementation of the deep learning models and interpretability techniques in this study used several key libraries. TensorFlow (2.13.0) and its Keras API were used for building and training the neural network models,

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

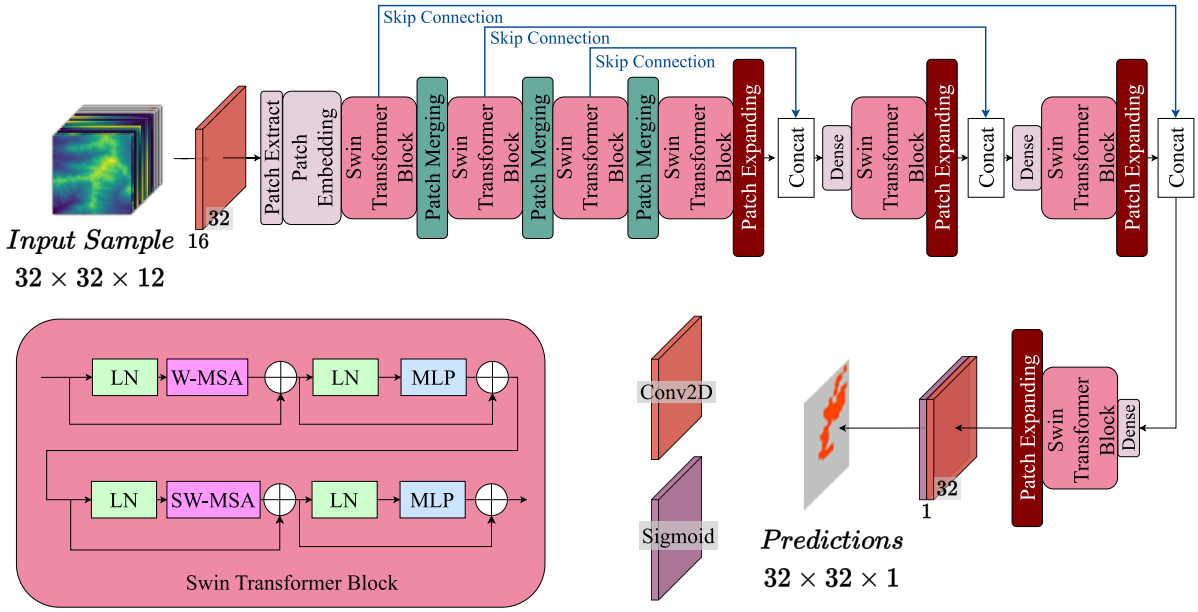


Figure 6. Detailed structure of Transformer-based Swin-UNet, featuring skip connections that extend from the encoder to the decoder, represented by blue lines in the illustration.

including the Autoencoder, ResNet, UNet, and Vision Transformer (ViT). These tools provided robust capabilities for creating complex architectures and training them efficiently. NumPy (1.25.0) handled numerical computations and array operations, essential for data preprocessing and manipulation tasks.

2.3. Evaluation Metrics

Given the nature of our task, which involves predicting wildfires from remote sensing data, we have chosen metrics that can provide a detailed understanding of model performance, particularly in the context of class imbalances and the critical importance of certain predictions. Our chosen metrics include:

1. GFLOPs (Giga Floating Point Operations Per Second): This metric measures the computational complexity of the model in terms of the number of floating-point operations performed per second. GFLOPs provide a quantitative assessment of how demanding a model is in terms of computational resources, which is crucial for understanding its feasibility for deployment in real-time systems or on devices with limited processing power. Models with high GFLOPs might be more accurate but could suffer from longer inference times and higher power consumption, which are critical factors in applications like wildfire prediction where timely response is essential.
2. Number of Parameters: This metric reflects the total count of trainable parameters within a model. A higher number of parameters generally indicates a more complex model that can capture more intricate patterns in the data. However, it also implies greater memory requirements and potential overfitting, especially in cases where data is scarce or noisy. Balancing the number of parameters is key to building efficient models that generalize well to new, unseen data without consuming excessive computational resources.
3. AUC-PR (Area Under the Curve—Precision-Recall): This metric is particularly beneficial for our study due to its sensitivity to class imbalances. Unlike the more commonly used Area Under the Curve—Receiver Operating Characteristics (AUC), AUC-PR focuses on the relationship between precision (the proportion of true positive results among all positive predictions) and recall (the proportion of true positive results detected among all relevant samples). This focus makes AUC-PR more informative for data sets with a significant imbalance between classes, as is often the case in wildfire prediction, where the presence of fire is a relatively rare event compared to its absence.
4. Precision and Recall with Masked Class: To further customize our evaluation to the specific challenges of our data set, we have implemented custom versions of precision and recall metrics that specifically exclude uncertain labels, which are represented as "−1" in our data set. This approach ensures that our evaluation metrics

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

only consider relevant classes for wildfire prediction, enhancing the focus on accurately predicting fire presence or absence without being disturbed by masked regions in the data.

5. AUC: While AUC is less sensitive to class imbalance than AUC-PR, it remains a valuable metric for evaluating overall model performance. AUC measures the ability of the model to distinguish between classes across all thresholds, providing a comprehensive overview of model effectiveness.
6. Confusion Matrix: This metric visualizes the performance of a classification model by displaying the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) in a matrix format. It helps in understanding the types of errors made by the model and the classes that are most often misclassified. Including a confusion matrix in our evaluation allows for a deeper analysis of how well the model performs specifically in distinguishing between the presence and absence of wildfires, considering the actual and predicted classifications.
7. Pearson Correlation Coefficient: This statistical measure assesses the linear relationship between two features, providing insights into the degree of correlation between the predicted and actual values. A higher Pearson correlation coefficient indicates a stronger direct linear relationship. In this study, the Pearson correlation is calculated based on the raw data values of the corresponding features.
8. Structural Similarity Index Measure (SSIM): This metric evaluates the visual impact of changes between two images, which is particularly useful in tasks that involve image processing or comparison. SSIM provides a more perceptual-based measure that compares the structural information in the images, offering a different perspective than purely pixel-based differences. In this study, the SSIM is calculated based on the raw data values of the corresponding features.

By leveraging these metrics, we aim to gain a thorough understanding of our models' predictive capabilities, with a particular emphasis on their ability to recognize and accurately predict wildfire occurrences.

2.4. Interpretability Techniques

In this study, we incorporated SHAP, Grad-CAM, and IG to achieve a better understanding of the implemented models. Before investigating the results of those techniques, it's essential to understand the underlying logic and mathematical principles of them.

2.4.1. SHAP

Shapley values, rooted in game theory and originally introduced by Shapley in 1953 (Shapley, 1953), provides the mathematical foundation for quantifying the importance of features through the average marginal contribution of each feature to the model's output. This ensures equitable attribution by considering all feature combinations. Later, Lundberg and Lee (2017) extended these concepts to machine learning explainability, specifically adapting SHAP for complex models, including tree-based algorithms like XGBoost and deep learning architectures, thus establishing SHAP explanations as a modern interpretability tool (Lellep et al., 2022; Z. Li, 2022).

The Shapley value which quantifies average contribution for a feature i in a model is mathematically defined as:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S)) \tag{1}$$

where $\phi_i(v)$ is the Shapley value for feature i. Here, S represents a subset of features excluding feature i. N is the set of all features in the model, with $|N|$ representing the total number of features. Additionally, $|S|$ indicates the number of features in subset S. The expression $v(S \cup \{i\}) - v(S)$ calculates the marginal contribution of feature i when added to the subset S. Function $v(S)$ is the prediction of the model using the features in subset S, whereas $v(S \cup \{i\})$ is the prediction using features in S along with feature i. The coefficient $\frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!}$ serves as a weighting factor that accounts for the number of permutations of feature subsets, ensuring a fair distribution of contribution among all features. The Shapley value therefore refers to the average contribution of each feature to the predictive outcome of a model, and also allowing us to understand the importance of a feature. In this study, we utilized Gradient Explainer from the SHAP library to compute Shapley values efficiently, leveraging its compatibility with diverse deep learning architectures, including convolutional neural networks and transformers, as well as its adaptability to high-dimensional inputs such as our data set. To establish a baseline model and

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

further analyze the SHAP differences between tree-based models and deep learning models, we also employed XGBoost, imported from the "xgboost" library, on the same data set with Tree Explainer.

2.4.2. Grad-CAM

Grad-CAM, crucial for enhancing the interpretability of deep learning models, especially CNNs, was introduced by Selvaraju et al. (2017). It provides a heatmap of influential regions within images for specific class predictions. This paper seeks to leverage Grad-CAM to visualize and compare the attention mechanisms inherent in the different approaches we have explored, namely Autoencoder, ResNet, UNet, and Transformer-based Swin-UNet. Consider the selected feature maps $\{A^k\}_{k=1}^{K}$ (K kernels from the final convolutional layer of a classifier), and let y^c be the logit corresponding to a specific class c. Grad-CAM computes the mean of the gradients of y^c across all N pixels (identified by coordinates u, v) within each feature map A^k, resulting in a weight w_k^c that denote its importance. The heatmap

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k w_k^c A^k\right) \quad \text{with} \quad w_k^c = \frac{1}{N}\sum_{u,v}\frac{\partial y^c}{\partial A^k_{u,v}} \tag{2}$$

is then produced by summing the feature maps with the respective weights, followed by the application of a pixel-wise ReLU operation to set negative values to zero. This process ensures that only regions positively influencing the class c decision are emphasized. While a classification network outputs a single class probability distribution for each input image $\mathbf{x}$, a segmentation model (such as our wildfire segmentation approaches) assigns logits $y^c_{i,j}$ to every pixel $x_{i,j}$ for each class c. Thus, Vinogradova et al. (2020) modified the original equation by substituting y^c with $\sum_{(i,j)\in M} y^c_{i,j}$, where M denotes the set of pixel indices of interest within the output mask:

$$L^c_{\text{SEG-Grad-CAM}} = \text{ReLU}\left(\sum_k w_k^c A^k\right) \quad \text{with} \quad w_k^c = \frac{1}{N}\sum_{u,v}\frac{\partial \sum_{(i,j)\in M} y^c_{i,j}}{\partial A^k_{u,v}} \tag{3}$$

SEG-Grad-CAM is well-suited for fire segmentation tasks because it processes logits from all fire-affected regions in an image, allowing for detailed spatial localization of fire areas. Unlike classification models that produce a single logit indicating the presence of fire, SEG-Grad-CAM can generate comprehensive activation maps that highlight each specific area where fire is detected. This method significantly advances our comprehension of how models interpret remote sensing data at the pixel level, a key for environmental monitoring. Additionally, we employ SEG-Grad-CAM to approximate the visualization of our data set's original 12 features, directing our attention to the first Conv2D layer. This strategy captures low-level features, ensuring a closer resemblance to the original data's visual characteristics through a channel-specific analysis.

This approach generates individual heatmaps for each of the 16 channels in the first convolutional layer of each model, as well as a combined heatmap that aggregates the contributions of all channels. This dual visualization provides a comprehensive understanding of how each channel and the collective set of channels influence the model's predictions. Our method, integrating individual and combined attention heatmaps, facilitates a thorough understanding of the model's information processing and decision-making, thereby identifying crucial feature regions for accurate predictions.

2.4.3. Integrated Gradients

Grad-CAM effectively highlights features with distinct visual patterns, such as "Elevation", "Drought", "Vegetation", "Population density", and "Previous Fire Mask" (PFM), by mapping their contributions onto heatmaps. However, for abstract or less visually discernible features, such as "Min temp" and "Humidity", Grad-CAM struggles to provide meaningful insights due to the lack of clear spatial characteristics. In these cases, IG complements Grad-CAM by providing quantitative measures of feature importance(Sundararajan et al., 2017), enabling a more comprehensive understanding of both visually apparent and abstract features. Integrated Gradients quantifies the contribution of input features by calculating the gradient of the model's output relative to each input feature along a linear path from a baseline to the actual input. The attribution for each input pixel value x_i against a baseline pixel value x_i' is calculated by:

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1
GFlops and Number of Parameters for 4 Models

	Autoencoder	ResNet	UNet	Transformer-based Swin-UNet
GFlops	0.038	0.452	0.024	1.146
Number of Parameters (M)	0.046	31.332	0.353	25.995

$$IG_i = (x_i - x_i') \times \int_{\lambda=0}^{1} \frac{\partial F(x' + \lambda \times (x - x'))}{\partial x_i} d\lambda \tag{4}$$

where F represents the model and λ varies from 0 to 1. This integral effectively captures the accumulated gradient contributions along the path from the baseline to the actual input, proportionally scaled. For our experiments, we used a zero baseline ($x_i' = 0$) for all features, which is a common choice in XAI methods to represent the absence of input signals.

In wildfire segmentation from remote sensing data, IG enables precise analysis at the pixel level, identifying features crucial for fire prediction. Employing IG, we obtain a $12 \times 32 \times 32$ attribution array per sample, reflecting the contributions of 12 features. By aggregating the values within each feature's matrix, we derive 12 comprehensive scores, quantifying the overall importance of each feature in the model's predictions. However, the analysis remained challenging since scores hardly reflect the degree of feature importance. Therefore, we calculated the positive contribution ratio (PCR) for each feature, revealing the proportion of each feature's positive contribution compared to others. For the i-th feature:

$$\mathrm{PCR}_i = \frac{\max(\gamma_i, 0)}{\sum_{i=1}^{12} \max(\gamma_i, 0)} \tag{5}$$

where γ_i represents the feature contribution of the i-th feature, for $i = 1, 2, \ldots, 12$. This method allows us to discern key features determining fire presence, enhancing our model's interpretability and guiding future model development. Integrating IG with Grad-CAM offers a thorough framework for analyzing feature contributions in wildfire prediction, enriching our understanding of the model's decision-making process.

2.5. Experimental Design

Our research methodology adopts a systematic experimental framework and examines the explainability and interpretability of four advanced models: Autoencoder, ResNet, UNet, and Transformer-based Swin-UNet. This framework is structured to ensure the reproducibility and reliability of models, and also to test the performance of the models under different training situations. In particular, we test the models under the following two variations:

1. Variation in Random Seed

To ensure the reproducibility and reliability of these models, an experiment will be conducted in which each model undergoes training four times. The training, validation, and test data sets are fixed, but different random seeds are selected for initialize each training session. This choice is designed to probe the models' performance consistency, effectively controlling for the variability introduced by stochastic initialization.

2. Variation in Data set Proportion

To investigate how the volume of training set impacts our models' performance, the validation set and test set were fixed, and each model will systematically be trained across a spectrum of training set proportions: 10%, 25%, 50%, 75%, and 100%. This tiered training approach is instrumental in evaluating the models' adaptability and performance across varying data availability. This experiment helps us distinguish models that are more adept at learning from limited data from those that require larger data sets to achieve optimal performance, offering insights into real-world wildfire prediction scenarios given varying data availability.

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2
Performance Metrics Evaluated on the Test Set, Averaged Over 3 Random Seeds

Model	AUC ± error rate	AUC-PR ± error rate	Precision ± error rate	Recall ± error rate
Autoencoder	0.8457 ± 0.64%	0.2338 ± 2.18%	0.3418 ± 3.83%	0.2551 ± 13.14%
ResNet	0.8274 ± 2.39%	0.1980 ± 3.28%	0.2985 ± 10.62%	0.2368 ± 35.98%
UNet	0.8463 ± 2.13%	0.2739 ± 3.14%	0.3464 ± 2.42%	**0.3859** ± 2.72%
Transformer-based Swin-UNet	**0.8637** ± 0.44%	**0.2803** ± 3.46%	**0.3686** ± 1.19%	0.3470 ± 5.53%

Note. The error rates reflect the variability of model performance across different initializations, with larger error rates indicating greater instability relative to the model's own metrics.

3. Results and Discussions

This section analyzes performance metrics for various models under distinct conditions, using Tables 2 and 3 to present metrics from training with different random seeds and data set volumes. The bold values indicate the best performance for each metric across different data fractions. Figure 7 visualizes the impact of initial seed variability on AUC-PR, parameters, and GFlops, while Figure 8 illustrates how data set volumes affect models performance. In Tables 4 and 5, "Predict Fire Mask" shows the prediction result according to each model. These visuals offer clear insights into the differences between models and the influences of seed settings and data size.

3.1. Fundamental Performance Metrics

AUC and AUC-PR are critical indicators of model performance, in Figures 7 and 8, the Transformer-based Swin-UNet and UNet models generally outperform Autoencoder and ResNet in both AUC and AUC-PR scores, indicating their superior ability to correctly predict wildfire events. The closeness in AUC scores among all models suggests they are generally comparable in distinguishing between fire and non-fire areas. However, the larger variance in AUC-PR scores highlights that the Transformer-based Swin-UNet and UNet are particularly

Table 3
Training Results Evaluated on Test Set for Different Fractions of Data Set

Fraction	Model	AUC	AUC-PR	Precision	Recall
10%	Autoencoder	0.8477	0.2321	0.2788	0.3903
	ResNet	0.7146	0.0578	0.2471	0.0004
	UNet	0.8340	**0.2643**	**0.3381**	0.3875
	Transformer-based Swin-UNet	**0.8520**	0.2593	0.3159	**0.4125**
25%	Autoencoder	0.8527	0.2111	0.2858	0.3367
	ResNet	0.8105	0.1819	0.2662	0.2755
	UNet	0.8467	0.2621	0.3211	0.3972
	Transformer-based Swin-UNet	**0.8699**	**0.2814**	**0.3412**	**0.4069**
50%	Autoencoder	0.8343	0.2247	0.3120	0.3142
	ResNet	0.8454	0.2049	0.2862	0.2835
	UNet	0.8469	0.2688	0.3389	0.3745
	Transformer-based Swin-UNet	**0.8726**	**0.2921**	**0.3644**	**0.3917**
75%	Autoencoder	0.8608	0.2433	0.3359	0.2900
	ResNet	0.8387	0.1867	0.2718	0.2658
	UNet	0.8586	**0.2807**	**0.3604**	**0.3752**
	Transformer-based Swin-UNet	**0.8696**	0.2720	0.3544	0.3493
100%	Autoencoder	0.8475	0.2278	0.3172	0.2883
	ResNet	0.8273	0.1870	0.3006	0.1644
	UNet	0.8337	0.2832	**0.3584**	**0.4040**
	Transformer-based Swin-UNet	**0.8707**	**0.2887**	0.3540	0.3880

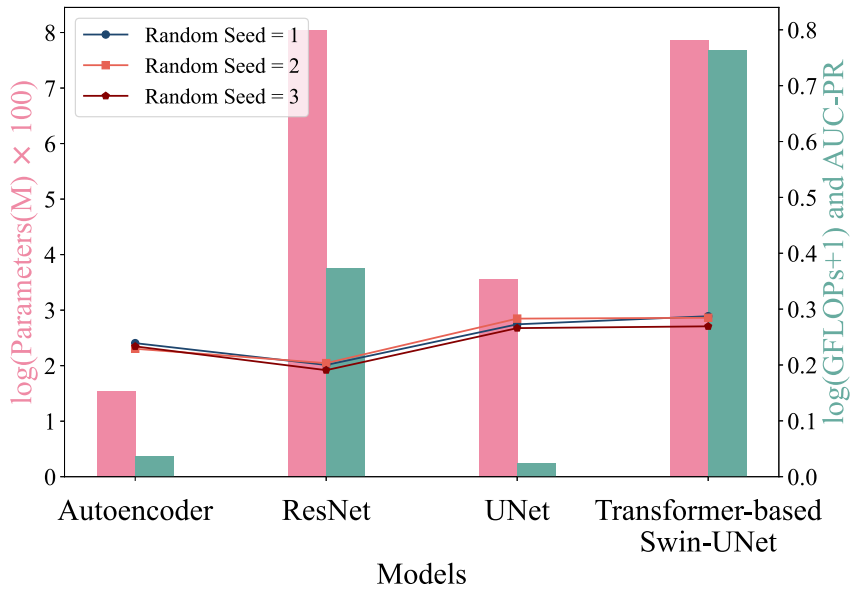


Figure 7. Comparison of models trained with different random seed. Each line represents the AUC-PR for each model with different initialization.

better at balancing precision and recall, crucial for imbalance scenarios such as wildfire. To further analyze the practical implications of these performance metrics, we examine the confusion matrices of each model along with Precision and Recall.

This section focuses specifically on the data set imbalance and its impact on model performance. Due to a significantly higher number of negative (i.e., without fire) samples compared to positive (i.e., with fire) samples, this imbalance is crucial for evaluating model performance. As shown in Figure 9, all models exhibit a significant discrepancy with TN far outnumbering TP, indicating a data set bias toward "No Fire" instances. Transformer-based Swin-UNet and UNet models perform better in identifying TP instances, reducing the occurrence of FN, which explains why these two models typically have higher Recall in Tables 2 and 3. As shown in the "Predict

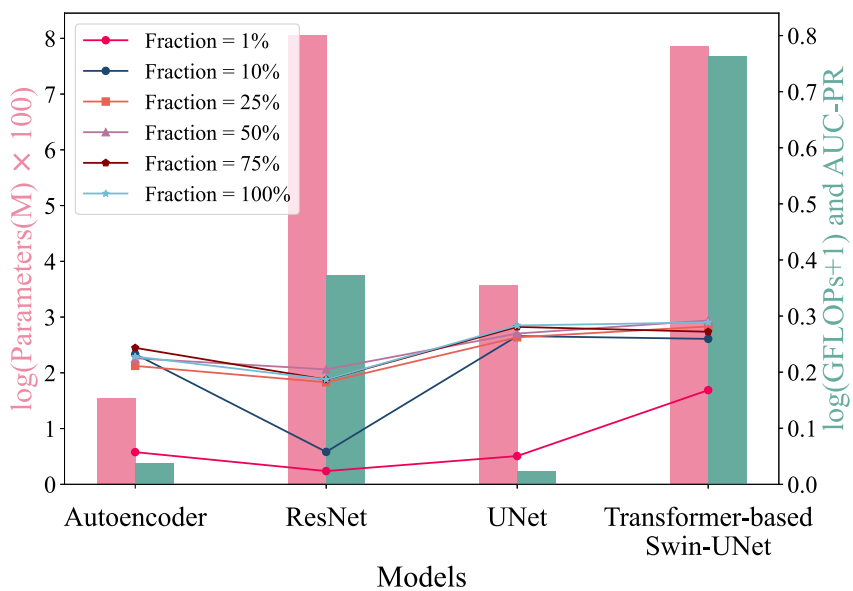


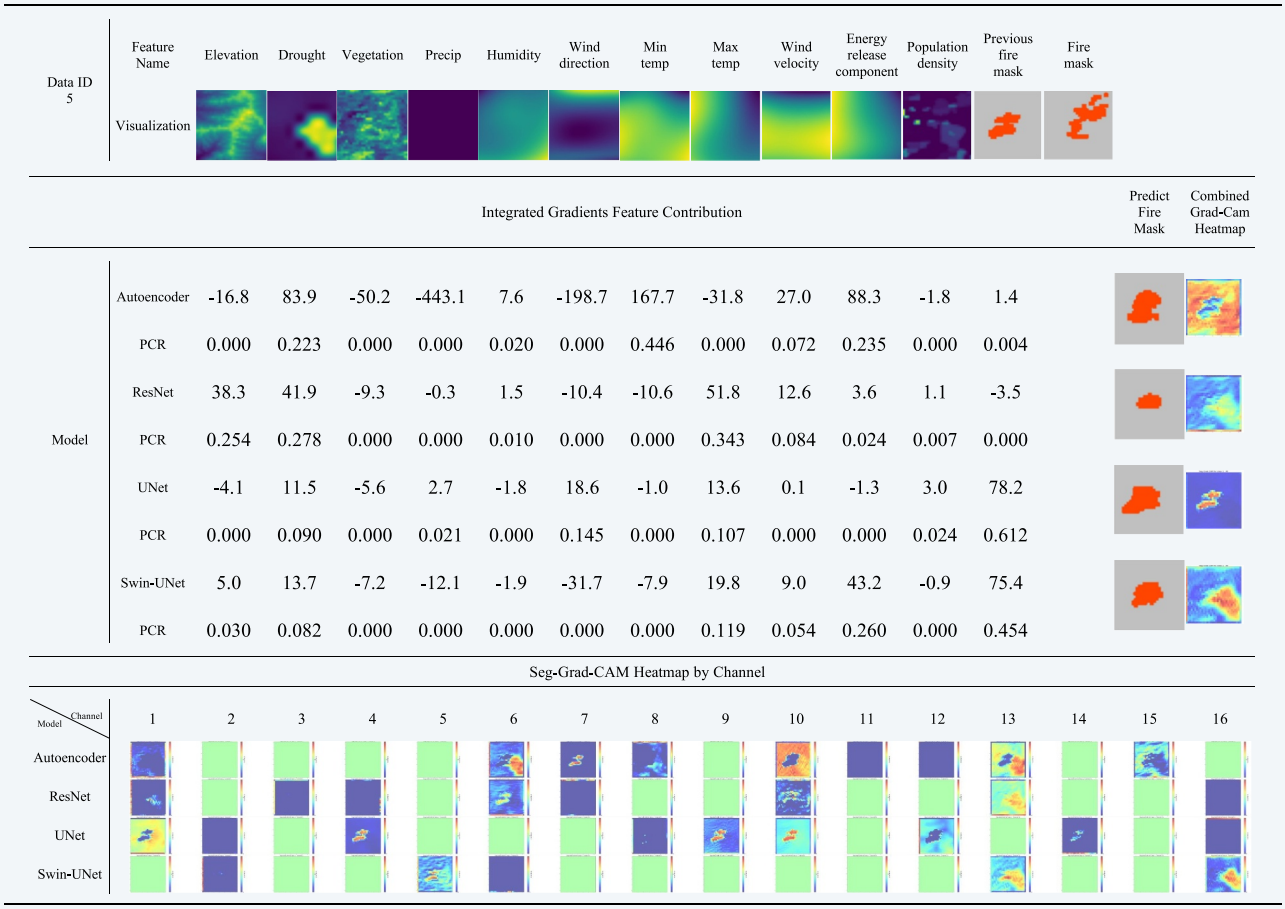
Figure 8. Comparison of models trained with different fraction of the data set. Each line represents the AUC-PR for each model with different training volume.

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10.1029/2024JH000409

Table 4

Analysis of the 5th Sample in the Data Set[a]



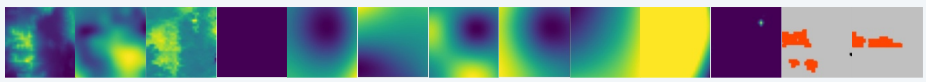
[a]In the analysis of SEG-Grad-CAM attention heatmaps by channel (16 subfigures representing the 16 channels of the first convolutional layer for each model), the heatmaps illustrate the recognition of feature patterns by each individual channel. It is evident that models such as Autoencoder, ResNet, UNet, and Transformer-based Swin-UNet can capture PFM, "Drought", and "Vegetation"to a certain extent. However, in the combined attention heatmap, the Autoencoder and ResNet models exhibit a lack of the PFM feature, which could be attributed to the models' attention being diverted by other features. In IG analysis, the feature contribution of PFM is significantly higher in UNet and Transformer-based Swin-UNet, in contrast to Autoencoder and ResNet, where it is considerably lower. Less critical features, such as 'Drought', disproportionately capture attention in the Autoencoder and ResNet models.

Fire Mask" part of Tables 4, 5, A2, A5, and A7, UNet and Transformer-based Swin-UNet usually have a wider range of predictions compared to the Autoencoder and ResNet. This broader range helps UNet and Transformer-based Swin-UNet to identify more TP instances, which in turn reduces the occurrence of FN. Specifically, the wider prediction range of UNet and Transformer-based Swin-UNet allows these models to capture more subtle variations in the data that might be missed by the Autoencoder and ResNet. In contrast, the Autoencoder and ResNet tend to have a more conservative prediction range, which often leads to higher FN rates. This conservatism means that FN instances in these models are more likely to appear near the boundary of the TP region, where the prediction confidence is lower. As a result, UNet and Transformer-based Swin-UNet typically have higher Recall, as shown in Tables 2 and 3.

On the other hand, the UNet and Transformer-based Swin-UNet might be more sensitive to positive samples in handling imbalanced data sets, but they also produce more FP, as illustrated in Figure 9, where UNet and Transformer-based Swin-UNet show more FP than the other two models. However, the other two models' performance in recognizing TP is significantly worse than that of UNet and Transformer-based Swin-UNet, resulting in generally higher Precision for UNet and Transformer-based Swin-UNet in Tables 2 and 3. Comparing UNet and Transformer-based Swin-UNet directly, UNet produces 0.4% more FP than Transformer-based Swin-UNet, and Transformer-based Swin-UNet produces 0.1% more FN than UNet, while Transformer-based Swin-UNet's

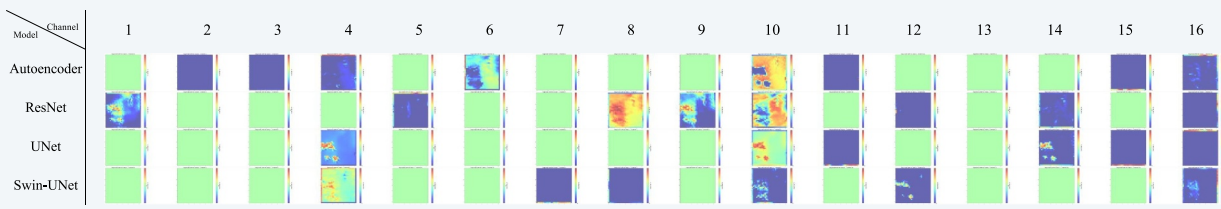
29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 5
Analysis of the 15th Sample in the Data Set[b]

Data ID 15	Feature Name	Elevation	Drought	Vegetation	Precip	Humidity	Wind direction	Min temp	Max temp	Wind velocity	Energy release component	Population density	Previous fire mask	Fire mask
	Visualization													

		Integrated Gradients Feature Contribution												Predict Fire Mask	Combined Grad-Cam Heatmap
Model	Autoencoder	82.3	87.3	-59.4	-71.4	-42.1	-14.4	7.8	7.3	-2.2	-385.5	-3.3	38.3		
	PCR	0.369	0.392	0.000	0.000	0.000	0.000	0.035	0.033	0.000	0.000	0.000	0.172		
	ResNet	39.2	-23.4	-51.1	18.9	47.9	4.0	-44.3	-0.2	0.1	151.0	-4.7	4.4		
	PCR	0.148	0.000	0.000	0.071	0.180	0.015	0.000	0.000	0.000	0.569	0.000	0.016		
	UNet	31.3	-12.5	-5.6	5.3	12.0	0.5	6.1	6.6	0.6	10.0	0.8	41.2		
	PCR	0.274	0.000	0.000	0.047	0.105	0.004	0.054	0.057	0.005	0.087	0.007	0.360		
	Swin-UNet	63.8	36.6	-61.8	17.0	-30.1	3.2	10.7	15.3	2.6	-12.2	-2.5	141.2		
	PCR	0.220	0.126	0.000	0.059	0.000	0.011	0.037	0.053	0.009	0.000	0.000	0.486		

Seg-Grad-CAM Heatmap by Channel



[b]In the analysis of SEG-Grad-CAM attention heatmaps by channel (16 subfigures representing the 16 channels of the first convolutional layer for each model), the heatmaps illustrate the recognition of feature patterns by each individual channel. It is observed that Autoencoder barely captures the features of the PFM, while ResNet, UNet, and Transformer-based Swin-UNet manage to detect PFM, "Drought", and "Vegetation" to a certain extent. In the combined attention heatmap, the absence of PFM features in Autoencoder suggests that the model's focus may have been diverted by other features. In contrast, ResNet, UNet, and Transformer-based Swin-UNet effectively highlight the PFM along with features such as 'Vegetation'. IG analysis reveals a high feature contribution for PFM in UNet and Transformer-based Swin-UNet, whereas it is notably low in Autoencoder and ResNet. Less critical features like "Drought" disproportionately occupy the attention of Autoencoder and ResNet.

TP is only 0.1% less than UNet's. This accounts for UNet typically having a higher Recall and Transformer-based Swin-UNet a higher Precision as noted in the same tables. These differences highlight distinct trade-off strategies between precision and recall, making each model suitable for different scenarios.

The UNet model, with its high Recall, is extremely suited for wildfire monitoring scenarios in California, where missing a fire could have serious consequences. In vast and remote forest areas, high-risk meteorological regions, and near critical infrastructure, it is crucial to detect every potential fire promptly. UNet is capable of capturing as many real fire events as possible, although this may accompany a higher rate of false alarms. In these scenarios, however, the ability to quickly identify and respond to potential fires outweighs the importance of reducing false alarms. The Transformer-based model, with its high Precision, is ideally suited for wildfire monitoring scenarios in California where minimizing false alarms is crucial. For example, in ecologically sensitive areas and high-value asset protection zones, frequent false alarms could lead to unnecessary interventions and resource wastage. In these cases, Transformer-based model's high precision ensures that only high-confidence fire alarms are triggered, thus optimizing resource use and minimizing environmental or community impacts. In a larger wildfire monitoring network, UNet can serve as a high-sensitivity layer, while Transformer-based Swin-UNet can act as a high-precision initial screening layer, allowing the system to capture all potential fires promptly and

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

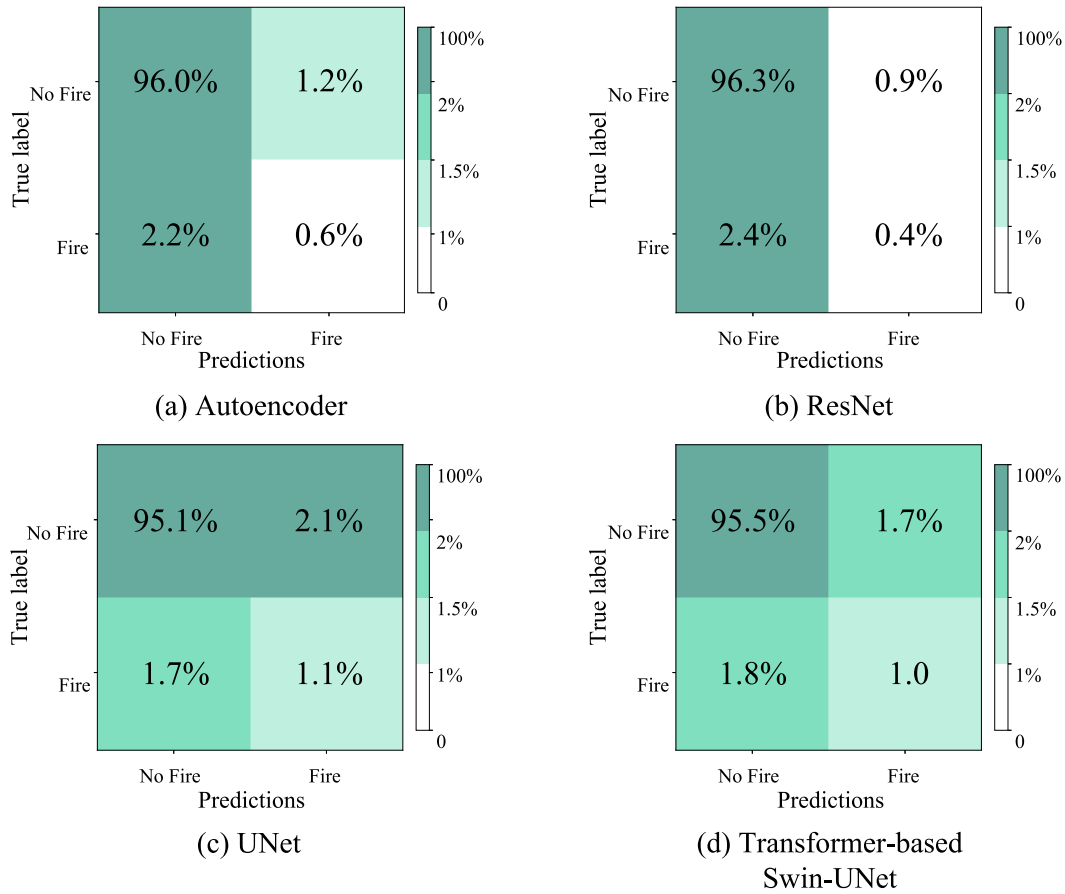


Figure 9. Confusion matrices for four models. The color scale is nonlinear due to the imbalance in the data set.

reduce false alarms through Transformer-based model's precise filtering. This setup leverages the complementary strengths of both models: UNet ensures no potential fire is missed, and Transformer-based Swin-UNet ensures that only high-confidence alarms are processed further. This multilayered design not only enhances the comprehensiveness of fire detection but also optimizes the effective use of resources, making it especially suitable for large-scale wildfire monitoring networks with extensive areas and precise resource management needs.

3.2. Efficiency of Model Architectures

Table 1 and the bars in Figures 7 and 8 indicate a subtle correlation between complexity and performance. Models like UNet, despite their lower computational load and parameter count, exhibited performance on par with or surpassing more complex counterparts such as ResNet and Transformer-based Swin-UNet. This challenges the notion that higher complexity equals better performance, showing that efficiency-optimized models can achieve significant accuracy.

In order to gain a deeper understanding of the underlying reasons for these results, an extensive analysis was conducted on the architectural features of the models. It was observed that Transformer-based Swin-UNet and UNet demonstrate superior performance compared to Autoencoder and ResNet. Take the average results from Table 2 as an example, the AUC-PR of UNet and Transformer-based Swin-UNet are 19.89% and 17.15% higher than that of Autoencoder, and 38.33% and 41.57% higher than that of ResNet. This improvement can be partially attributed to their distinctive approaches to implementing skip connections. Unlike ResNet and Autoencoder, which only use skip connections within blocks, UNet and Transformer-based Swin-UNet employ them across the encoder and decoder, improving the integration of deep and surface-level features for detecting complex wildfire spread patterns. In UNet and Transformer-based Swin-UNet, skip connections help bridge the gap between the input data and the final prediction layer, allowing for the preservation of crucial information that might otherwise

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

be lost in deeper layers. In contrast, Autoencoder and ResNet incorporate skip connections only within their blocks to enhance information flow. This architectural choice in UNet and Transformer-based Swin-UNet is instrumental in enhancing the models' ability to accurately predict wildfires, as it ensures that both high-level and low-level features are effectively synthesized to facilitate the prediction process. The success of UNet and Transformer-based Swin-UNet can thus be partially attributed to their architecture's inherent capacity to maintain a rich, integrated feature set across the network, highlighting the importance of such structural considerations in model design.

In comparing Transformer-based Swin-UNet and UNet, Transformer-based Swin-UNet has significantly more parameters and GFlops but only slightly better predictive performance. Transformer-based Swin-UNet's complexity arises from its self-attention mechanism, which suits high-dimensional data like images but increases computational complexity due to extensive matrix operations. Additionally, Transformer-based Swin-UNet's multi-head attention mechanism boosts parameter count and complexity, demanding more computational resources. Despite Transformer-based Swin-UNet's significant computational enhancements, its modest performance gains do not offer clear advantages in all scenarios. For example, in wildfire monitoring where quick and accurate predictions are crucial, UNet is preferable for real-time prediction and emergency measures due to its efficiency. Transformer-based Swin-UNet, however, may be more suited for post-event analysis and firefighting strategy formulation due to its higher precision and longer computation time.

3.3. Robustness Across Initialization Variants

As presented in Figure 7, each curve represents the AUC-PR of different models on the test data set under the same random seed. The close clustering of points for each model (marked in four colors) suggests that the AUC-PR of each model varies little under different random seeds. Testing models with different random seeds is a common method to evaluate robustness because it helps to assess the consistency of the model's performance despite variations in the initial conditions.

The error rates of each metric are shown in Table 2. These results indicate that Autoencoder, UNet, and Transformer-based Swin-UNet demonstrate relatively lower Precision and Recall error rates, which means their performance remains more consistent under different initial conditions introduced by the random seeds. This consistency is a key indicator of robustness. In contrast, the high variability of ResNet in Recall (error rate up to 35.98%) and Precision (error rate up to 10.62%) suggests that this model is particularly sensitive to variations in data set features or label distributions caused by different initializations, which could lead to unstable performance in practical applications. Therefore, while considering models for this scenario, where stability and reliability are crucial, the sensitivity of ResNet to initial conditions must be carefully evaluated.

3.4. Influence of Data Volume on Training

As presented in Figure 8, each curve represents the AUC-PR of different models trained with the same data volume. As the volume of data increases, both UNet and Transformer-based Swin-UNet models show a trend of improvement in performance metrics such as AUC and AUC-PR. In particular, the Transformer-based Swin-UNet shows a more significant performance increment as the amount of data increases, indicating that more training data helps the UNet and Transformer-based Swin-UNet learn and generalize better. However, the improvement in performance is not linear in all cases. Especially for the ResNet model, at certain data volumes, performance improvement reaches a plateau or shows fluctuation, possibly reflecting the model's limited adaptability to data complexity (Jafar & Lee, 2021). Meanwhile, the Autoencoder's performance does not show a trend of improvement. This may be due to its architecture being insufficient to capture the complex features and patterns in the data, particularly with increased data volume and complexity, since it has the lowest number of parameters (Alzubaidi et al., 2021).

In essence, this section evaluates model performances under varying conditions, demonstrating that UNet and Transformer-based Swin-UNet outperform ResNet and Autoencoder in key metrics. Transformer-based Swin-UNet, in particular, stands out for its balanced accuracy and false alarm rate. The analysis also reveals the models' robustness and the effects of data volume. While increased data generally benefits learning, ResNet and Autoencoder do not follow this trend, highlighting the importance of model selection based on specific task requirements and data characteristics.

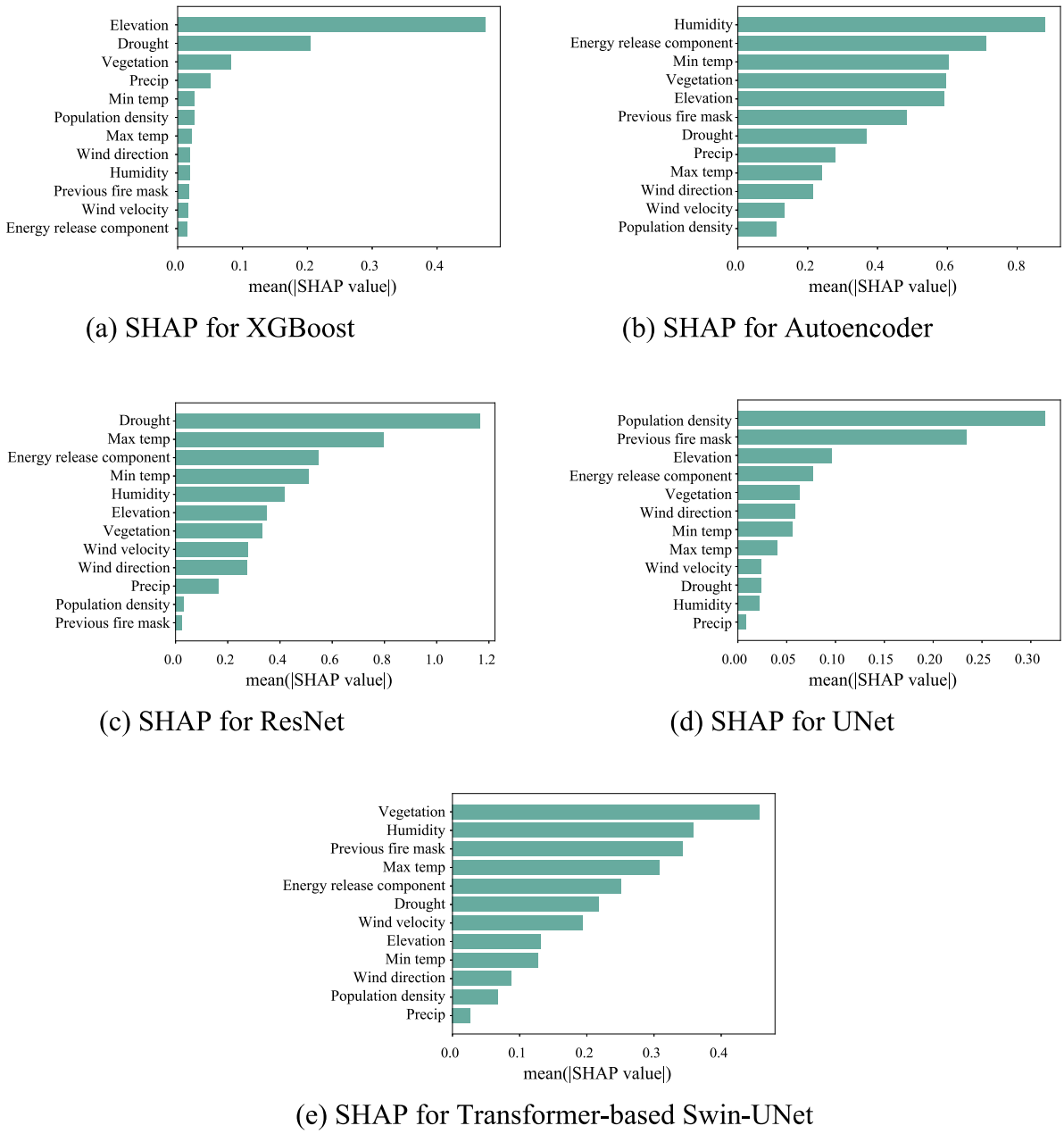


Figure 10. Mean absolute SHAP value (The average impact on model output magnitude, calculated as the mean of the absolute SHAP values for each feature across the entire data set).

4. Interpretability Analysis

4.1. SHAP

From the experiments conducted by Huot (Huot et al., 2022b), it was discovered that removing the PFM from the training data resulted in the poorest performance of the model, thereby triggering our interest in the PFM. A clear pattern is that the degree to which models prioritize the PFM feature is closely linked to their performance in predicting wildfire spread. In the UNet and Transformer-based Swin-UNet, where PFM ranks high (second in Figure 10d and third in Figure 10e, respectively), we observe a correlation with these models' superior performance (AUC-PR values of 0.2739 and 0.2803, respectively). This underscores the ability of UNet and Transformer-based Swin-UNet to recognize and utilize this spatial feature, which is directly related to wildfire spread. The Autoencoder demonstrates a moderate level of performance (AUC-PR = 0.2338), with PFM's

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

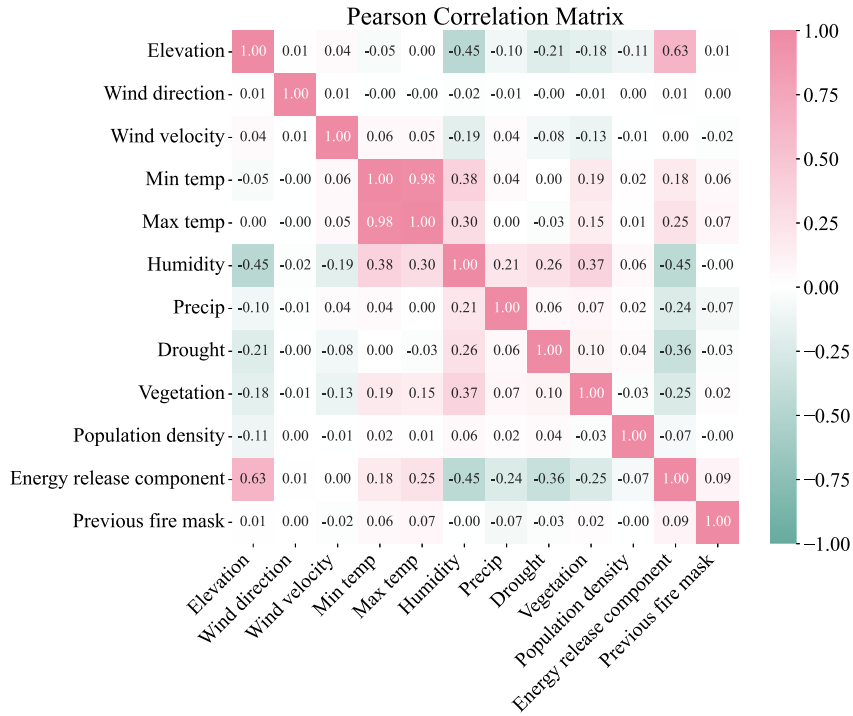


Figure 11. This matrix visualizes the pairwise Pearson correlation coefficients between features. Shades of pink represent positive correlations approaching 1, while shades of green indicate negative correlations nearing −1. The color gradient provides a clear visual indication of the strength and direction of correlations.

moderate ranking in Figure 10b. Conversely, in the XGBoost and ResNet models, where PFM ranks low (third from last in Figure 10a and last in Figure 10c, respectively), there is an alignment with their lower prediction performance (AUC-PR values of 0.06 and 0.1980), suggesting these models may not fully leverage the PFM feature. This further affirms the viewpoint that PFM is a crucial feature for wildfire spread prediction.

Except for PFM, we can also find other points worth analyzing from Figure 10. UNet ranked "Population density" higher than other models (1st place in Figure 10d), which indicates its strong reliance on this feature for prediction. Notably, despite Population density's low importance in Transformer-based Swin-UNet (11th place in Figure 10e), its AUC-PR score is slightly higher than that of UNet. Transformer-based Swin-UNet prioritizes "Vegetation" and "Humidity" as the most important features, while UNet assigns them medium level of importance. This distinction may explain the slight advantage of Transformer-based Swin-UNet over UNet in wildfire spread prediction tasks. Transformer-based Swin-UNet's self-attention mechanism allows for a deep understanding of complex relationships between features and spatial context, especially the impact of "Vegetation" on the spread of fire. This capability may enable Transformer-based Swin-UNet to outperform in capturing these critical dynamics. UNet outperforms in image segmentation and capturing spatial data, yet Transformer-based Swin-UNet surpasses it by integrating climate factors more flexibly through its self-attention mechanism. Aside from the previously discussed features, the remaining ones are considered variably across models. Their impact on wildfire prediction is nuanced, with each contributing to understanding the complex interplay of factors influencing fire spread. However, their relative importance and the way they are integrated vary, reflecting the diverse approaches of models to balance spatial, climatic, and human factors in predicting wildfire dynamics (Figure 11).

4.2. Grad-CAM

Following the macroscopic SHAP analysis, we identified several intriguing findings that warrant further investigation. For example, in some models, the importance of "Population density" was notably high, while others exhibited an imbalance in attention to crucial features. To investigate deeper, we employed IG and Grad-CAM for a more detailed analysis of individual samples. From Table 4 to A7, we present a comprehensive

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

comparison of IG feature contribution and Grad-CAM attention heatmaps, along with visualizations of the corresponding features. By integrating these analytical tools, we hope to gain a thorough understanding of the feature contributions and model behavior specific to this data set.

However, before taking IG analysis into consideration, we can already find some patterns from SEG-Grad-CAM: Firstly, from Tables 4, 5, A5, and A7, the Autoencoder's lack of attention toward the PFM is evidenced by predominantly blue regions, indicating a deficiency in capturing PFM features. In contrast, UNet and Transformer-based Swin-UNet exhibit a heightened focus on PFM, as shown by yellow or red areas, demonstrating their superior ability to identify and emphasize PFM. This difference underscores the capability of UNet and Transformer-based Swin-UNet to effectively recognize and prioritize PFM features, which are critical for model performance. Secondly, UNet and Transformer-based Swin-UNet's ability to concentrate on PFM does not detract from their attention to other important features such as "Vegetation" and "Drought". This is evident from the detailed contours for these features in the "Combined Grad-CAM Heatmap" of Table 4, 5, A1, A2, A5, which are more pronounced than those observed in the Autoencoder and ResNet models. However, in most cases, Transformer-based Swin-UNet is more effective than UNet at capturing and retaining information on variables such as "Vegetation" and "Drought". Although ResNet can also capture these features to some extent, its performance is hindered by an overemphasis on PFM, which may lead to an imbalance in capturing other critical climate variables. Transformer-based Swin-UNet's superiority may be due to its integration of a global attention mechanism and skip connections in its advanced architecture. The global attention mechanism allows Transformer-based Swin-UNet to dynamically identify and focus on the most important information throughout the entire image, which is particularly beneficial for processing variables with complex spatial and temporal distribution features, such as "Vegetation" and "Drought".

Additionally, skip connections ensure the effective transfer and preservation of important detail information between the deep layers of the model, which helps in more accurately reconstructing these critical climate data features during the decoding phase. Therefore, Transformer-based Swin-UNet's performance in capturing and retaining information on variables like "Vegetation" and "Drought" seems surpass that of Autoencoder, ResNet, and UNet. Furthermore, the analysis highlights that models shift their attention toward the "Population density" when PFM is missing. In such cases, as observed in Table A1, A3, A4, and A6, the "Combined Grad-CAM Heatmap" of all four models shows contours of the "Population density", which typically does not appear in cases where the PFM is present normally. Lastly, "Elevation" is typically an important feature, but it is difficult to discern its outline in the "Combined Grad-CAM Heatmap". By analyzing the features in Tables 5, A1–A7, we find that "Elevation" often shows a high similarity with the feature "Vegetation", which is also evidenced by the SSIM evaluation. As illustrated in Figure 12, the SSIM score of "Elevation" against "Vegetation", highlighted in a yellow box, is 0.93, indicating a high visual similarity between the two features. Thus, we speculate that this high degree of similarity might lead to difficulties in distinguishing between these two features in the heatmap generation. This could explain why "Elevation" is not prominently visible in the heatmap, as the model may blend its information with that of "Vegetation", thereby masking its unique visual identity.

Grad-CAM provides an intuitive visual understanding for features with easily recognizable shapes, such as "Elevation", "Drought", "Vegetation", and "Population density". However, it is less effective for features that share similarities and abstract features like "Humidity" and "Min temp". These limitations arise because similar features tend to blend together and abstract features lack distinct shapes, making them difficult to distinguish on heatmaps. Therefore, we used IG to help analyze these abstract features and quantitatively assess the contribution of each feature.

4.3. Integrated Gradients

Through IG analysis, we can not only better quantify abstract features that are difficult to distinguish in Grad-CAM heatmaps but also cross-validate the analyses from Grad-CAM and SHAP. In UNet and Transformer-based Swin-UNet, the PCR of the PFM feature is typically greater compared to other models (see Tables 4, 5, A1, A2, A5, and A7). This indicates that these architectures may place more emphasis on PFM. UNet and Transformer-based Swin-UNet also focus on features like "Drought" and "Elevation", but usually they don't surpass the crucial PFM (see Table 5, A1, and A7). This suggests a balanced recognition and utilization of various features, with PFM remaining crucial in their predictive decision-making. Notably, this emphasis on PFM aligns with the findings presented by Huot et al. (2022b), which underscore the critical role of historical fire information

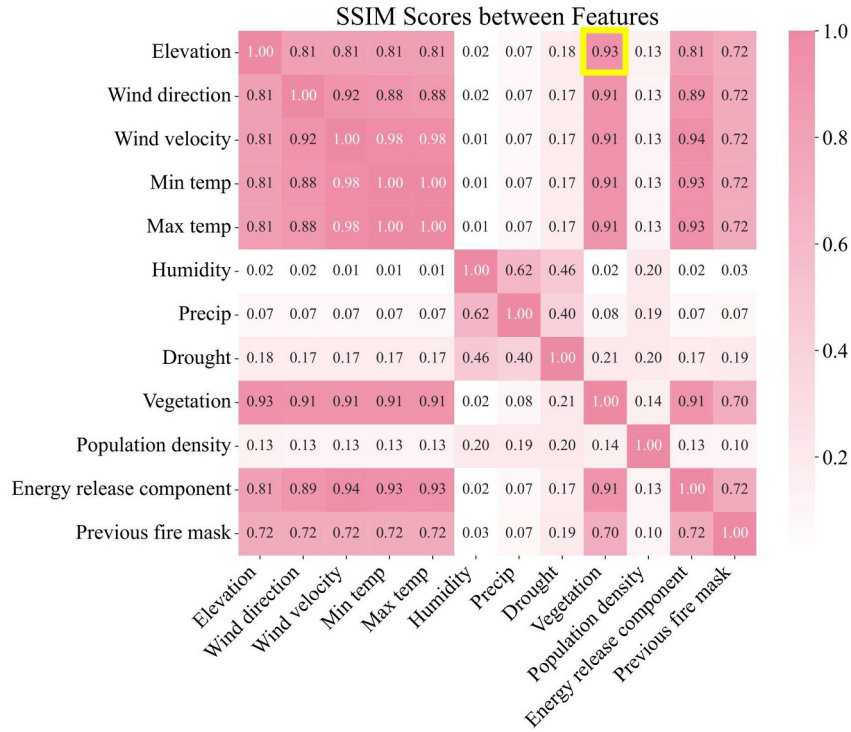


Figure 12. Structural Similarity Index Measure (SSIM) scores between features. This matrix displays the SSIM scores, which measure the similarity between different features. The score for "Elevation" against "Vegetation" is highlighted in a yellow box. Shades of pink indicate a score of 1, representing perfect similarity, while white represent a score of 0, indicating no similarity.

in predicting future wildfire spread. The convergence of insights from SHAP, IG, and Grad-CAM analyses further strengthens our understanding of feature importance. For instance, the high feature importance of the PFM in SHAP analysis for UNet is mirrored in the IG and Grad-CAM analyses, where its significance is visually and quantitatively confirmed.

In Autoencoder and ResNet, when "Drought" or "Vegetation" is significant, they often overshadow PFM in terms of PCR (see Tables 4, 5, A1, A2, A4, and A5). This implies a diverted focus from other features, leading to less effective accommodation by the Autoencoder. On the contrary, Transformer-based Swin-UNet and UNet achieve a better balance, meaning that while focusing on other features such as "Drought", "Vegetation", and "Elevation", these models still maintain sufficient attention on PFM. This point also corresponds with the findings in Grad-CAM, which highlights a crucial aspect of model tuning. The ability of UNet and Transformer-based Swin-UNet to maintain balanced attention to critical features contributes significantly to their superior performance. This balance ensures that the models do not overly prioritize certain features at the expense of others, facilitating a more accurate and comprehensive understanding of the factors driving the spread of wildfires. UNet and Transformer-based Swin-UNet achieve this through skip connections, which help preserve important details by combining shallow and deep features. These findings provide valuable insights for future model selection and design. Incorporating skip connections can enhance a model's ability to retain essential features across layers. This strategy facilitates the development of models that offer a more accurate and comprehensive understanding of multifaceted environmental phenomena, such as wildfire spread.

These congruence across interpretability tools not only validates the models' reliance on critical features but also exemplifies how SHAP, IG, and Grad-CAM can collectively enhance our interpretability framework. Their combined application enables a comprehensive and detailed exploration of feature contributions, facilitating more transparent and explainable machine learning models in environmental science.

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5. Conclusion

This study began a comprehensive exploration of deep learning models' capabilities in predicting wildfires using remote sensing data, emphasizing the importance of model interpretability. Rather than arbitrarily selecting models, our choice was aligned with mainstream machine learning methods that predominantly employ either CNN or Transformer-based architectures. This alignment is important as these architectures represent the cutting edge in handling the complex spatial data typical of remote sensing applications. Through this, we analyzed the performances and interpretability of prominent representatives from these categories—specifically, Autoencoder, ResNet, and UNet from CNNs, and Transformer-based Swin-UNet from Transformer architectures. Our findings provide detailed insights into their predictive performances and the critical role of interpretability in their application to wildfire prediction.

Firstly, our findings indicate that the UNet and Transformer-based model exhibit superior predictive accuracy compared to the Autoencoder and ResNet models. This superior performance can be attributed, in part, to their different implementations of skip connections, which facilitate effective feature transmission across network layers, enhancing the model's ability to learn and generalize from complex spatial data, and maintain focus on critical features such as PFM. Secondly, in the comparison between Transformer-based Swin-UNet and UNet, Transformer-based Swin-UNet slightly outperforms due to its adoption of a global attention mechanism and multi-head attention mechanism. These mechanisms enable Transformer-based Swin-UNet to more comprehensively capture and analyze high-dimensional data, such as climate features. This explains why Transformer-based Swin-UNet performs better than UNet in certain scenarios, as it can integrate and interpret large-scale spatial data better. Lastly, the application of interpretability tools such as SHAP, IG, and Grad-CAM has provided deeper insights into the decision-making processes of these models. These tools highlight the importance of the PFM feature and the need for balanced feature representation to enhance prediction accuracy. Besides, Grad-CAM analysis shows that Transformer-based Swin-UNet's attention mechanism not only focuses on crucial features like PFM but also effectively captures other significant variables such as “Drought” and “Vegetation”. This capability allows Transformer-based Swin-UNet to more accurately predict wildfire spread and development, demonstrating its unique strengths in feature balance and information extraction.

Through our comparison of models, we also provide guidance for model selection to accommodate different situations. Firstly, UNet and Transformer-based Swin-UNet offer distinctive trade-offs between precision and recall, making each suitable for specific scenarios. One reason for the difference is the implementation of the self-attention mechanism in Transformer-based Swin-UNet. The self-attention mechanism enables Swin UNet to more flexibly capture global dependencies when processing input data, helping the model to more accurately distinguish important features. This contributes to its superior performance in precision, thereby making Transformer-based Swin-UNet optimal for reducing false alarms in sensitive areas and avoiding unnecessary interventions. UNet, with its high recall, is ideal for critical areas where missing a fire could be catastrophic, despite a higher false alarm rate. Secondly, the complexity of the UNet and Transformer-based Swin-UNet models also makes them suitable for different scenarios. Transformer-based Swin-UNet has significantly more parameters and GFlops but only slightly better predictive performance, making UNet preferable for real-time prediction and emergency measures due to its efficiency in wildfire monitoring where quick and accurate predictions are crucial. Transformer-based Swin-UNet, however, may be more suited for post-event analysis and firefighting strategy formulation due to its higher precision and longer computation time. Lastly, from our robustness experiments with the four models, we believe ResNet, due to its lack of robustness, should be carefully considered for this scenario, as stability and reliability are essential.

Future research could investigate refining model architectures to enhance both predictive accuracy and interpretability. Investigating hybrid models that take the strengths of UNet and Transformer-based models together may offer more robust predictive capabilities for wildfire forecasting. Further scrutiny of the role of skip connections in processing this particular data set, alongside trials with other models incorporating such features, is warranted. Moreover, the slight advantage of Transformer-based Swin-UNet over UNet—whether it stems from the transformer structure's advantages or merely a greater number of parameters—merits further investigation. Additionally, enhancing the use of interpretability tools in model assessment could promote increased trust and transparency in deploying deep learning for environmental surveillance and disaster response.

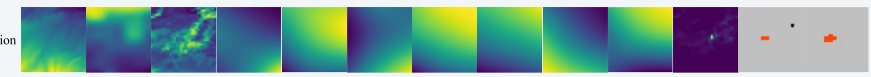
29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Appendix A: Tables for IG and Grad-CAM Analysis

Table A1.

Table A1
Analysis of the 17th Sample in the Data Set[c]

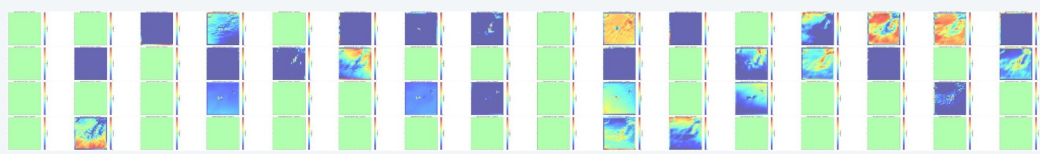
Data ID 17

Feature Name	Elevation	Drought	Vegetation	Precip	Humidity	Wind direction	Min temp	Max temp	Wind velocity	Energy release component	Population density	Previous fire mask	Fire mask
Visualization													

Integrated Gradients Feature Contribution

Model	Elevation	Drought	Vegetation	Precip	Humidity	Wind direction	Min temp	Max temp	Wind velocity	Energy release component	Population density	Previous fire mask	Predict Fire Mask	Combined Grad-Cam Heatmap
Autoencoder	-446.9	10.7	105.6	-50.7	-64.4	-20.2	-68.6	60.7	7.1	-12.3	-8.8	10.8		
PCR	0.000	0.055	0.542	0.000	0.000	0.000	0.000	0.311	0.037	0.000	0.000	0.055		
ResNet	46.7	-5.0	13.1	0.0	1.1	31.6	1.3	12.8	-11.0	3.3	4.1	0.1		
PCR	0.410	0.000	0.115	0.000	0.009	0.277	0.011	0.112	0.000	0.029	0.036	0.001		
UNet	-2.1	-1.6	0.0	-1.2	1.8	0.8	-2.1	-0.5	13.0	0.1	2.5	9.1		
PCR	0.000	0.000	0.000	0.000	0.065	0.028	0.000	0.000	0.477	0.002	0.092	0.335		
Swin-UNet	-1.1	1.2	1.7	-1.5	3.6	-2.2	-5.5	-1.3	-1.4	-0.6	1.3	2.7		
PCR	0.000	0.111	0.164	0.000	0.341	0.000	0.000	0.000	0.000	0.000	0.124	0.260		

Seg-Grad-CAM Heatmap by Channel

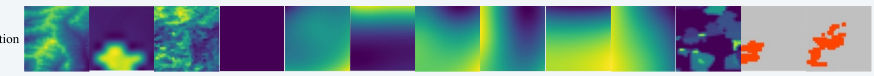
Model \ Channel	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Autoencoder																
ResNet																
UNet																
Swin-UNet																

[c]In the analysis of SEG-Grad-CAM attention heatmaps by channel, it appears that possibly due to the small size of the PFM, Autoencoder, ResNet, and Transformer-based Swin-UNet have almost failed to capture its features, with UNet being the exception. It is evident that all four models have successfully identified features related to "Drought" and "Vegetation". In the combined attention heatmap, UNet clearly captures the PFM, while the other three models barely show the PFM shape, likely due to its small size. IG analysis indicates that the feature contribution of PFM remains significantly high in UNet and Transformer-based Swin-UNet, but notably low in Autoencoder and ResNet. Less critical features, such as "Vegetation", have disproportionately drawn the attention of Autoencoder and ResNet.

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

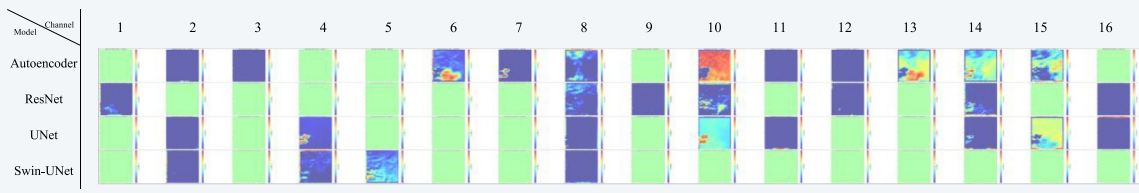
Table A2.

Table A2
Analysis of the 21st Sample in the Data Set[d]

Data ID 21	Feature Name	Elevation	Drought	Vegetation	Precip	Humidity	Wind direction	Min temp	Max temp	Wind velocity	Energy release component	Population density	Previous fire mask	Fire mask
	Visualization													

		Integrated Gradients Feature Contribution												Predict Fire Mask	Combined Grad-Cam Heatmap
Model	Autoencoder	82.3	87.3	-59.4	-71.4	-42.1	-14.4	7.8	7.3	-2.2	-385.5	-3.3	38.3		
	PCR	0.369	0.392	0.000	0.000	0.000	0.000	0.035	0.033	0.000	0.000	0.000	0.172		
	ResNet	39.2	-23.4	-51.1	0.000	47.9	4.0	-44.3	-0.2	0.1	151.0	-4.7	4.4		
	PCR	0.159	0.000	0.000	0.000	0.194	0.016	0.000	0.000	0.000	0.612	0.000	0.018		
	UNet	31.3	-12.5	-5.6	5.3	12.0	0.5	6.1	6.6	0.6	10.0	0.8	41.2		
	PCR	0.274	0.000	0.000	0.047	0.105	0.004	0.054	0.057	0.005	0.087	0.007	0.360		
	Swin-UNet	5.0	13.7	-7.2	-12.1	-1.9	-31.7	-7.9	19.8	9.0	43.2	-0.9	75.4		
	PCR	0.030	0.082	0.000	0.000	0.000	0.000	0.000	0.119	0.054	0.260	0.000	0.454		

Seg-Grad-CAM Heatmap by Channel



[d]In the SEG-Grad-CAM attention heatmaps by channel, it is observed that Autoencoder, ResNet, UNet, and Transformer-based Swin-UNet are capable of capturing features like the PFM, "Drought", and "Vegetation" to varying degrees. In the combined attention heatmap, Autoencoder shows a notable absence of PFM features, and ResNet displays these features faintly, which may be due to the models' attention being diverted by other features. Conversely, UNet and Transformer-based Swin-UNet outperform in highlighting both PFM and "Vegetation" features prominently. IG analysis further reveals that the feature contribution of PFM is significantly higher in UNet and Transformer-based Swin-UNet, whereas it remains quite low in Autoencoder and ResNet. Less critical features such as "Drought" occupy a substantial amount of attention in Autoencoder and ResNet models.

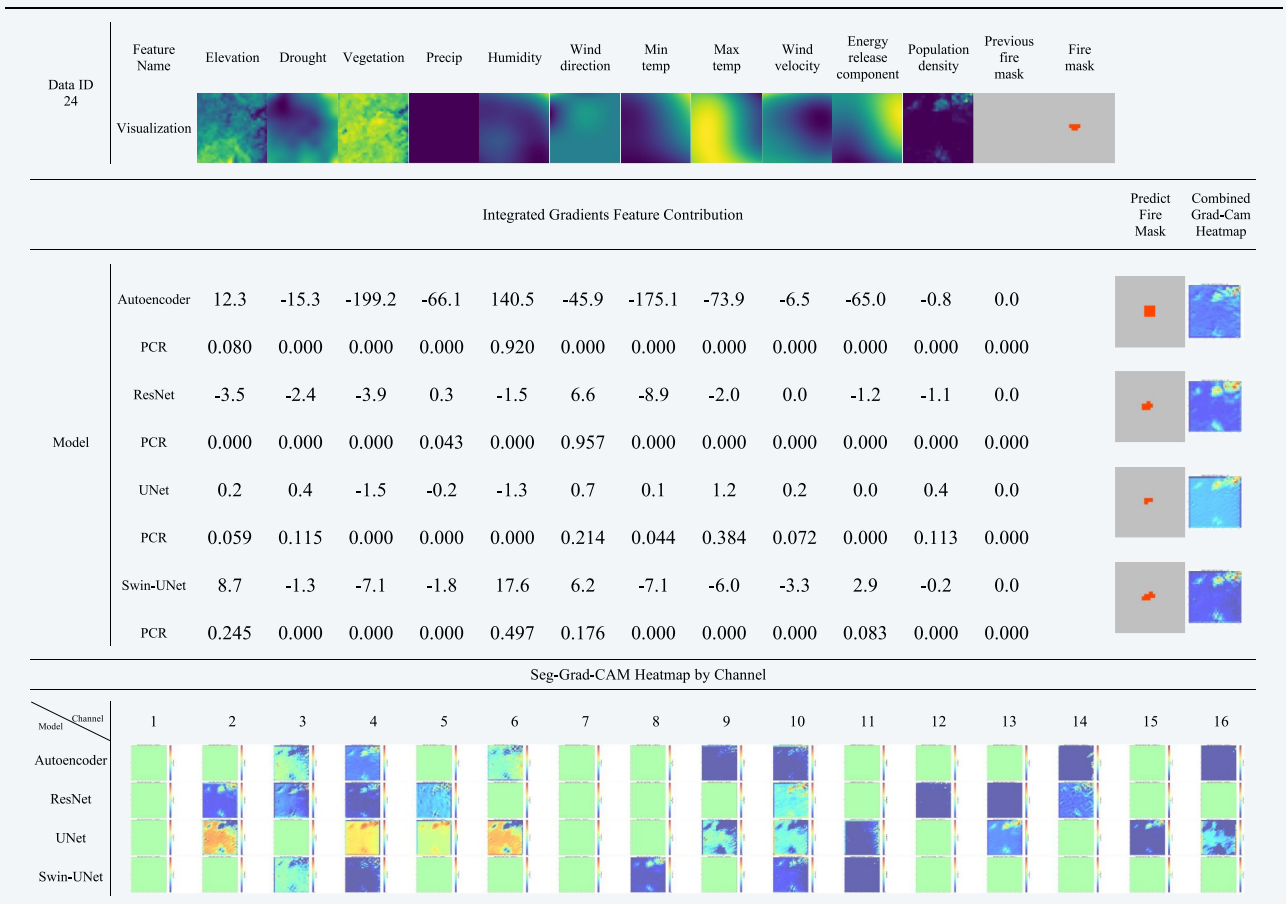
29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10.1029/2024JH000409

Table A3.

Table A3

Analysis of the 24th Sample in the Data Set[e]

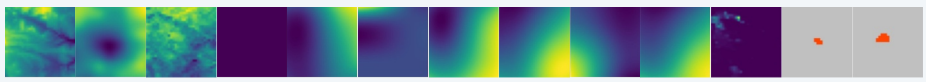


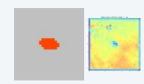
[e]Analysis of the 24th data. In the analysis of SEG-Grad-CAM attention heatmaps by channel, the absence of PFM is attributed to the lack of PFM data. It is observed that Autoencoder, ResNet, UNet, and Transformer-based Swin-UNet are all capable of detecting features like "Drought" and "Vegetation" to some extent. Notably, "Population density" feature, which was absent in most of the samples, is now clear. In the combined attention heatmap, the primary feature captured by all four models is the "Population density". According to the IG analysis, the feature contribution of PFM is zero across all models, which is consistent with the absence of PFM data in the data set.

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

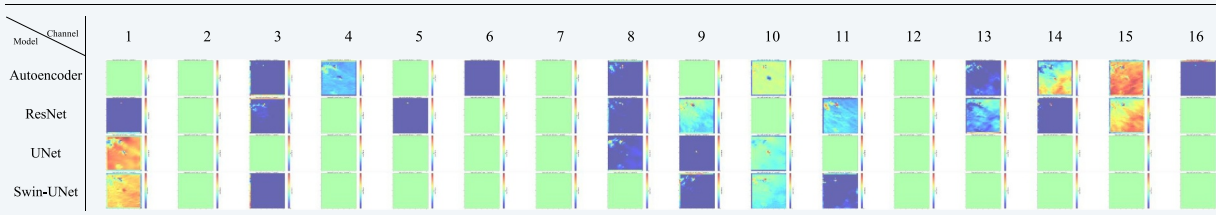
Table A4.

Table A4
Analysis of the 25th Sample in the Data Set[f]

Data ID 25	Feature Name	Elevation	Drought	Vegetation	Precip	Humidity	Wind direction	Min temp	Max temp	Wind velocity	Energy release component	Population density	Previous fire mask	Fire mask
	Visualization													

Integrated Gradients Feature Contribution													Predict Fire Mask	Combined Grad-Cam Heatmap
Model	Autoencoder	-367.5	110.1	14.3	-135.0	3.6	-102.5	-26.1	53.5	2.5	-14.0	6.0	-11.2	
	PCR	0.000	0.579	0.075	0.000	0.019	0.000	0.000	0.282	0.013	0.000	0.032	0.000	
	ResNet	103.3	-26.7	18.2	2.3	-32.0	6.4	-3.6	9.1	-24.4	-10.0	-18.0	2.3	
	PCR	0.730	0.000	0.128	0.016	0.000	0.045	0.000	0.064	0.000	0.000	0.000	0.017	
	UNet	-6.2	-1.4	0.7	-3.1	13.9	-2.8	-1.4	1.8	-5.9	1.4	18.5	6.0	
	PCR	0.000	0.000	0.016	0.000	0.328	0.000	0.000	0.042	0.000	0.033	0.439	0.142	
	Swin-UNet	-3.2	10.7	0.3	-1.8	3.7	-5.2	-1.2	-3.3	-2.1	-0.6	3.0	1.2	
	PCR	0.000	0.566	0.018	0.000	0.196	0.000	0.000	0.000	0.000	0.000	0.158	0.062	

Seg-Grad-CAM Heatmap by Channel



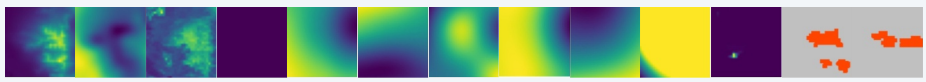
[f]In the analysis of SEG-Grad-CAM attention heatmaps by channel, it is observed that perhaps due to the small size of the PFM, Autoencoder and ResNet almost fail to capture its features, while UNet and Transformer-based Swin-UNet can capture it. All four models successfully captured "Drought" and "Vegetation". In the combined attention heatmap, UNet and Transformer-based Swin-UNet exhibit some shape of PFM, but the other two models, especially Autoencoder, show a clear absence of these features, probably attributed to the small size of PFM. IG analysis indicates that the feature contribution of PFM is not high in UNet and Transformer-based Swin-UNet, and remains low in Autoencoder and ResNet as well.

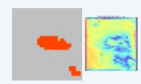
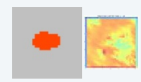
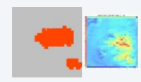
29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table A5.

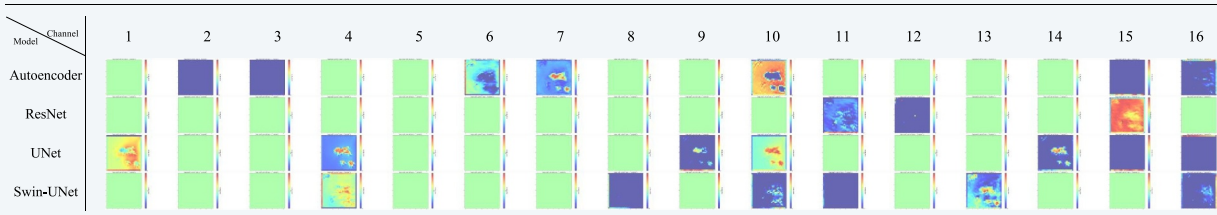
Table A5

Analysis of the 29th Sample in the Data Set[g]

Data ID 29	Feature Name	Elevation	Drought	Vegetation	Precip	Humidity	Wind direction	Min temp	Max temp	Wind velocity	Energy release component	Population density	Previous fire mask	Fire mask
	Visualization													

		Integrated Gradients Feature Contribution												Predict Fire Mask	Combined Grad-Cam Heatmap
Model	Autoencoder	37.1	61.9	-29.8	-145.3	-21.3	-10.6	14.3	-11.6	8.9	-291.0	-3.8	12.0		
	PCR	0.276	0.461	0.000	0.000	0.000	0.000	0.107	0.000	0.066	0.000	0.000	0.090		
	ResNet	-38.6	19.8	-4.3	-6.6	-55.4	-12.0	19.8	67.9	-1.1	130.6	0.2	6.3		
	PCR	0.000	0.081	0.000	0.000	0.000	0.000	0.081	0.277	0.000	0.534	0.001	0.026		
	UNet	10.7	1.6	-12.2	2.5	-19.5	-0.2	-2.6	0.0	0.7	31.7	0.7	63.4		
	PCR	0.096	0.015	0.000	0.022	0.000	0.000	0.000	0.000	0.006	0.285	0.006	0.570		
	Swin-UNet	20.8	-2.0	0.4	4.1	-10.6	0.4	-9.7	43.9	0.1	-11.2	2.4	87.3		
	PCR	0.131	0.000	0.003	0.026	0.000	0.002	0.000	0.275	0.000	0.000	0.015	0.548		

Seg-Grad-CAM Heatmap by Channel



[g]In the SEG-Grad-CAM attention heatmaps by channel, it is evident that Autoencoder, ResNet, UNet, and Transformer-based Swin-UNet are capable of capturing PFM, "Drought", and "Vegetation". In the combined attention heatmap, the PFM feature is missing in Autoencoder and not clear in ResNet, possibly due to these models' focus being diverted to other features. Conversely, UNet and Transformer-based Swin-UNet effectively highlight the PFM along with "Vegetation" features. IG analysis reveals that the feature contribution of PFM is significantly high in UNet and Transformer-based Swin-UNet, whereas it is notably low in Autoencoder and ResNet. This is in contrast to features like "Drought", which, despite being less critical, occupy a substantial amount of attention in Autoencoder and ResNet.

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

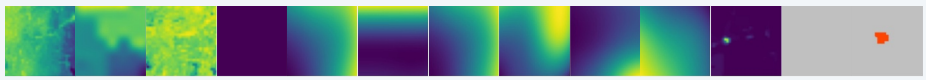
10.1029/2024JH000409

Table A6.

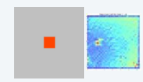
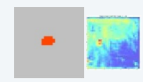
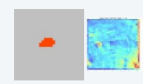
Table A6

Analysis of the 33rd Sample in the Data Set[h]

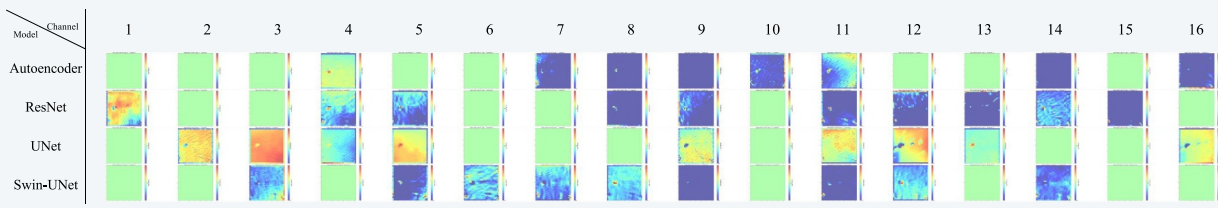
Data ID 33

Feature Name	Elevation	Drought	Vegetation	Precip	Humidity	Wind direction	Min temp	Max temp	Wind velocity	Energy release component	Population density	Previous fire mask	Fire mask
Visualization													

Integrated Gradients Feature Contribution

Model	Elevation	Drought	Vegetation	Precip	Humidity	Wind direction	Min temp	Max temp	Wind velocity	Energy release component	Population density	Previous fire mask	Predict Fire Mask	Combined Grad-Cam Heatmap
Autoencoder	-86.2	1.0	-67.4	-268.7	-45.2	-7.5	27.1	0.7	7.1	-49.2	-1.1	0.0		
PCR	0.000	0.027	0.000	0.000	0.000	0.000	0.755	0.020	0.198	0.000	0.000	0.000		
ResNet	-12.2	-0.6	-5.0	6.1	3.0	1.9	0.8	1.5	-0.5	-6.9	2.1	0.0		
PCR	0.000	0.000	0.000	0.395	0.194	0.126	0.050	0.096	0.000	0.000	0.139	0.000		
UNet	5.9	0.2	-1.4	-0.2	-1.4	0.2	1.7	0.0	0.3	-0.3	3.5	0.0		
PCR	0.500	0.016	0.000	0.000	0.000	0.018	0.144	0.000	0.027	0.000	0.294	0.000		
Swin-UNet	11.8	1.5	-7.3	-5.9	-7.9	-2.1	1.8	2.6	4.8	6.6	-0.3	0.0		
PCR	0.405	0.050	0.000	0.000	0.000	0.000	0.061	0.090	0.166	0.227	0.000	0.000		

Seg-Grad-CAM Heatmap by Channel



[h]In the SEG-Grad-CAM attention heatmaps by channel, the absence of PFM is due to the lack of such data in the data set. It is observed that Autoencoder, ResNet, UNet, and Transformer-based Swin-UNet are capable of capturing features like "Drought" and "Vegetation". Notably, the "Population density" feature, which was almost absent in other samples, is now more clear. In the combined attention heatmap, the primary feature captured by all four models is the "Population density". IG analysis reflects that, due to the absence of PFM data, the feature contribution of PFM is zero across all models. Meanwhile, the "Population density" plays a more significant role in ResNet and UNet than previously observed, indicating its increased relevance in the analysis of these models.

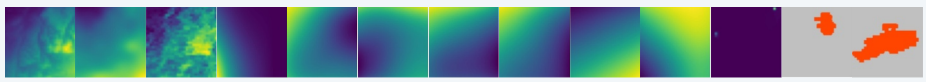
29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table A7.

Table A7

Analysis of the 125th Sample in the Data Set[i]

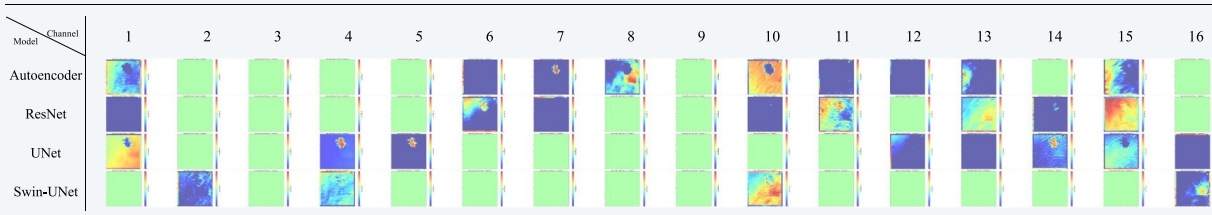
Data ID 125

Feature Name	Elevation	Drought	Vegetation	Precip	Humidity	Wind direction	Min temp	Max temp	Wind velocity	Energy release component	Population density	Previous fire mask	Fire mask
Visualization													

Integrated Gradients Feature Contribution

Model														Predict Fire Mask	Combined Grad-Cam Heatmap
	Autoencoder	-171.4	40.1	-58.3	-303.5	0.9	-137.8	137.0	-24.5	31.7	221.0	-13.9	-12.2		
	PCR	0.000	0.093	0.000	0.000	0.002	0.000	0.318	0.000	0.074	0.513	0.000	0.000		
	ResNet	31.8	28.7	5.5	5.0	-0.4	-36.3	35.4	56.1	-15.5	29.9	1.5	-9.6		
	PCR	0.164	0.148	0.028	0.026	0.000	0.000	0.182	0.289	0.000	0.154	0.008	0.000		
	UNet	-0.1	5.6	-13.3	1.8	-1.6	0.7	-2.4	6.2	-2.6	27.9	3.4	57.2		
	PCR	0.000	0.054	0.000	0.018	0.000	0.007	0.000	0.061	0.000	0.271	0.033	0.556		
	Swin-UNet	56.5	46.6	-8.5	-13.7	-6.3	-21.1	-27.0	25.2	23.1	36.8	-9.8	78.4		
	PCR	0.212	0.175	0.000	0.000	0.000	0.000	0.000	0.095	0.087	0.138	0.000	0.294		

Seg-Grad-CAM Heatmap by Channel



[i]In the analysis of SEG-Grad-CAM attention heatmaps by channel, we observed that Autoencoder, ResNet, UNet, and Transformer-based Swin-UNet are capable of capturing features of PFM, "Drought", and "Vegetation" to varying extents. However, in the combined attention heatmap, the PFM feature is missing in Autoencoder and not clear in Transformer-based Swin-UNet, which may be attributed to these models' attention being diverted to other features. IG analysis reveals that the feature contribution of PFM is significantly high in UNet and Transformer-based Swin-UNet, whereas it is zero in Autoencoder and ResNet, indicating no positive contribution from PFM in these two models. Features such as "Drought", which is less critical, attract more attention in Autoencoder and ResNet.

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Notation

β_i	i-th moment exponential decay rate
$\phi_i(v)$	Shapley value for feature i
S	a subset of features excluding feature i
N	the set of all features in the model
$\|N\|$	the total number of features
$\|S\|$	the number of features in subset S
$v(S \cup \{i\})$	the prediction using features in S along with feature i
$v(S)$	the prediction of the model using the features in subset S
$v(S \cup \{i\}) - v(S)$	the marginal contribution of feature i when added to the subset S
A^k	the k-th feature map in a convolutional layer
y^c	the logit corresponding to a specific class c
w_k^c	weights associated with the A^k
$y_{i,j}^c$	the logits to every pixel $x_{i,j}$ for the class$c$
M	the set of pixel indices of interest within the output mask
$w_{k,i,j}^c$	pixel-specific weights associated with the A^k
F	represents the model
x_i	a specific pixel value within one of the 12 feature channels
x_i'	The baseline input pixel value corresponding to x_i
λ	A scaling factor ranging from 0 to 1, used to interpolate between the baseline input x_i' and the actual input x_i
γ_i	the feature contribution of the i-th feature
PCR_i	positive contribution ratio for i-th feature
IG_i	the attribution for each input pixel value x_i against a baseline pixel value x_i'

Data Availability Statement

The multivariate dataset titled "Next Day Wildfire Spread" in this study is available at Kaggle (Huot et al., 2022a) with the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The codes used for implementing the model and the XAI tools for wildfire prediction and explanation, along with detailed documentation, are preserved on GitHub. These resources are accessible (Y. Zhou et al., 2024) under the MIT License, which permits free use, modification, and redistribution.

Acknowledgments
Sibo Cheng acknowledges the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-22-CPJ2-0143-01.

References

Abdollahi, A., & Pradhan, B. (2023). Explainable artificial intelligence (xai) for interpreting the contributing factors feed into the wildfire susceptibility prediction model. *The Science of the Total Environment*, *879*, 163004. https://doi.org/10.1016/j.scitotenv.2023.163004

Ahmad, K., Khan, M. S., Ahmed, F., Driss, M., Boulila, W., Alazeb, A., et al. (2023). Firexnet: An explainable ai-based tailored deep learning model for wildfire detection on resource-constrained devices. *Fire Ecology*, *19*(1), 54. https://doi.org/10.1186/s42408-023-00216-0

Al-Dabbagh, A. M., & Ilyas, M. (2023). Uni-temporal sentinel-2 imagery for wildfire detection using deep learning semantic segmentation models. *Geomatics, Natural Hazards and Risk*, *14*(1). https://doi.org/10.1080/19475705.2023.2196370

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8

Andela, N., Morton, D. C., Giglio, L., Paugam, R., Chen, Y., Hantson, S., et al. (2019). The global fire atlas of individual fire size, duration, speed and direction. *Earth System Science Data*, *11*(2), 529–552. https://doi.org/10.5194/essd-11-529-2019

Artés, T., Oom, D., De Rigo, D., Durrant, T. H., Maianti, P., Libertà, G., & San-Miguel-Ayanz, J. (2019). A global wildfire dataset for the analysis of fire regimes and fire behaviour. *Scientific Data*, *6*(1), 296. https://doi.org/10.1038/s41597-019-0312-2

Asensio, M., & Ferragut, L. (2002). On a wildland fire model with radiation. *International Journal for Numerical Methods in Engineering*, *54*(1), 137–157. https://doi.org/10.1002/nme.420

Bommer, P., Kretschmer, M., Hedström, A., Bareeva, D., & Höhne, M. M.-C. (2023). Finding the right xai method - A guide for the evaluation and ranking of explainable ai methods in climate science. *ArXiv*. https://doi.org/10.48550/arXiv.2303.00652

Bousfield, C., Lindenmayer, D., & Edwards, D. P. (2023). Substantial and increasing global losses of timber-producing forest due to wildfires. *Nature Geoscience*, *16*(2), 123–130. https://doi.org/10.1038/s41561-023-01323-y

Campbell, S., & Hossain, A. (2022). Application of remote sensing to study the potential impacts of 2020 wildfire events on the glaciers of mount baker, Washington. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLVI-M-2–2022*, 53–57. https://doi.org/10.5194/isprs-archives-xlvi-m-2-2022-53-2022

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv Preprint arXiv:2105.05537*. Retrieved from https://arxiv.org/abs/2105.05537

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature News*, *538*(7623), 20–23. https://doi.org/10.1038/538020a

Cheng, S., Guo, Y., & Arcucci, R. (2023). A generative model for surrogates of spatial-temporal wildfire nowcasting. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *7*(5), 1420–1430. https://doi.org/10.1109/tetci.2023.3298535

Cheng, S., Prentice, I. C., Huang, Y., Jin, Y., Guo, Y.-K., & Arcucci, R. (2022). Data-driven surrogate model with latent data assimilation: Application to wildfire forecasting. *Journal of Computational Physics*, *464*, 111302. https://doi.org/10.1016/j.jcp.2022.111302

Didan, K., & Barreto, A. (2018). *VIIRS/NPP vegetation indices 16-day L3 global 500 m SIN grid V001 (Technical Report)*. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/VIIRS/VNP13A1.001

Dikshit, A., & Pradhan, B. (2021). Interpretable and explainable ai (xai) model for spatial drought prediction. *The Science of the Total Environment*, *801*, 149797. https://doi.org/10.1016/j.scitotenv.2021.149797

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. Retrieved from https://arxiv.org/abs/2010.11929

Fan, D., Biswas, A., & Ahrens, J. P. (2024). Explainable ai integrated feature engineering for wildfire prediction.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., et al. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, *45*(2). https://doi.org/10.1029/2005RG000183

Giglio, L., & Justice, C. (2015). *MOD14A1 MODIS/terra thermal anomalies/fire daily L3 global 1 km SIN grid V006 (Technical Report)*. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD14A1.006

Girtsou, S., Apostolakis, A., Giannopoulos, G., & Kontoes, C. (2021). A machine learning methodology for next day wildfire prediction. In *2021 IEEE international geoscience and remote sensing symposium IGARSS* (pp. 8487–8490). https://doi.org/10.1109/IGARSS47720.2021.9554301

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Guo, Y., Wang, X., Shi, J., Sun, L., & Lan, X. (2023). Deep learning approaches for wildland fires using satellite data. *Remote Sensing*, *15*(5), 1192. https://doi.org/10.3390/rs15051192

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). https://doi.org/10.1109/cvpr.2016.90

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Hodges, J., & Lattimer, B. (2019). *Wildland fire spread modeling using convolutional neural networks* (pp. 1–28). Fire Technology. https://doi.org/10.1007/S10694-019-00846-4

Huot, F., Hu, R. L., Goyal, N., Sankar, T., Ihme, M., & Chen, Y.-F. (2022a). Next day wildfire spread. https://www.kaggle.com/datasets/fantineh/next-day-wildfire-spread

Huot, F., Hu, R. L., Goyal, N., Sankar, T., Ihme, M., & Chen, Y.-F. (2022b). Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data [Dataset]. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–13. https://doi.org/10.1109/TGRS.2022.3192974

Ivek, T., & Vlah, D. (2022). Reconstruction of incomplete wildfire data using deep generative models. *Extremes*, *26*(2), 1–21. https://doi.org/10.1007/s10687-022-00459-1

Jafar, A., & Lee, M. (2021). High-speed hyperparameter optimization for deep resnet models in image recognition. *Cluster Computing*, *26*(5), 2605–2613. https://doi.org/10.1007/s10586-021-03284-6

Jain, P., Coogan, S. C., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. *Environmental Reviews*, *28*(4), 478–505. https://doi.org/10.1139/er-2020-0019

Ji, Y., Wang, D., Li, Q., Liu, T., & Bai, Y. (2024). Global wildfire danger predictions based on deep learning taking into account static and dynamic variables. *Forests*, *15*(1), 216. https://doi.org/10.3390/f15010216

Khanmohammadi, S., Arashpour, M., Golafshani, E. M., Cruz, M. G., Rajabifard, A., & Bai, Y. (2022). Prediction of wildfire rate of spread in grasslands using machine learning methods. *Environmental Modelling and Software*, *156*, 105507. https://doi.org/10.1016/j.envsoft.2022.105507

Khryashchev, V., & Larionov, R. (2020). Wildfire segmentation on satellite images using deep learning. 2020 Moscow Workshop on Electronic and Networking Technologies (MWENT). *2020 Moscow Workshop on Electronic and Networking Technologies (MWENT)*, 1–5. https://doi.org/10.1109/MWENT47943.2020.9067475

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., et al. (2022). Wildfire danger prediction and understanding with deep learning. *Geophysical Research Letters*, *49*(17). https://doi.org/10.1029/2022GL099368

Laube, R., & Hamilton, H. J. (2021). Wildfire occurrence prediction using time series classification: A comparative study. In *2021 ieee international conference on big data (big data)* (pp. 4178–4182). https://doi.org/10.1109/BigData52589.2021.9671680

Laurent, P., Mouillot, F., Yue, C., Ciais, P., Moreno, M. V., & Nogueira, J. M. (2018). Fry, a global database of fire patch functional traits derived from space-borne burned area products. *Scientific Data*, *5*(1), 1–12. https://doi.org/10.1038/sdata.2018.132

Lellep, M., Prexl, J., Eckhardt, B., & Linkmann, M. (2022). Interpreted machine learning in fluid dynamics: Explaining relaminarisation events in wall-bounded shear flows. *Journal of Fluid Mechanics*, *942*, A2. https://doi.org/10.1017/jfm.2022.307

Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective kernel networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 510–519). https://doi.org/10.1109/CVPR.2019.00060

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost. *Computers, Environment and Urban Systems*, *96*, 101845. https://doi.org/10.1016/j.compenvurbsys.2022.101845

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Mamalakis, A., Barnes, E., & Ebert-Uphoff, I. (2022). Carefully choose the baseline: Lessons learned from applying xai attribution methods for regression tasks in geoscience. *ArXiv*. https://doi.org/10.48550/arXiv.2208.09473

Mell, W., Jenkins, M. A., Gould, J., & Cheney, P. (2007). A physics-based approach to modelling grassland fires. *International Journal of Wildland Fire*, *16*(1), 1–22. https://doi.org/10.1071/wf06002

Nolan, R., Anderson, L., Poulter, B., & Varner, J. (2022). Increasing threat of wildfires: The year 2020 in perspective: A global ecology and biogeography special issue. *Global Ecology and Biogeography*, *31*(9), 1655–1668. https://doi.org/10.1111/geb.13588

Pan, X., Ye, T., Xia, Z., Song, S., & Huang, G. (2023). Slide-transformer: Hierarchical vision transformer with local self-attention. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2082–2091). https://doi.org/10.1109/CVPR52729.2023.00207

Perry, G. (1998). Current approaches to modelling the spread of wildland fire: A review. *Progress in Physical Geography*, *22*(2), 222–245. https://doi.org/10.1177/030913339802200204

Pham, K., Ward, D., Rubio, S., Shin, D., Zlotikman, L., Ramirez, S., et al. (2022). California wildfire prediction using machine learning. In *2022 21st ieee international conference on machine learning and applications (icmla)* (pp. 525–530). https://doi.org/10.1109/ICMLA55696.2022.00086

Qayyum, F., Samee, N. A., Alabdulhafith, M., Aziz, A., & Hijjawi, M. (2024). Shapley-based interpretation of deep learning models for wildfire spread rate prediction. *Fire Ecology*, *20*(1), 8. https://doi.org/10.1186/s42408-023-00242-y

Rashkovetsky, D., Mauracher, F., Langer, M., & Schmitt, M. (2021). Wildfire detection from multisensor satellite imagery using deep semantic segmentation. *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 7001–7016. https://doi.org/10.1109/JSTARS.2021.3093625

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the ieee international conference on computer vision* (pp. 618–626).

Shapley, L. S. (1953). A value for n-person games. *Annals of Mathematical Studies*, *28*, 307–317.

Sullivan, A. L. (2009). Wildland surface fire spread modelling, 1990–2007. 1: Physical and quasi-physical models. *International Journal of Wildland Fire*, *18*(4), 349–368. https://doi.org/10.1071/wf06143

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*. Retrieved from https://arxiv.org/abs/1703.01365

Suwansrikham, P., & Singkhamfu, P. (2023). Performance evaluation of deep learning algorithm for forest fire detection. In *2023 joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT and NCON)* (pp. 244–248). https://doi.org/10.1109/ECTIDAMTNCON57770.2023.10139443

Tavakkoli Piralilou, S., Einali, G., Ghorbanzadeh, O., Nachappa, T. G., Gholamnia, K., Blaschke, T., & Ghamisi, P. (2022). A google earth engine approach for wildfire susceptibility prediction fusion with remote sensing data of different spatial resolutions. *Remote Sensing*, *14*(3), 672. https://doi.org/10.3390/rs14030672

Vinogradova, K., Dibrov, A., & Myers, G. (2020). Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34(10), pp. 13943–13944). https://doi.org/10.1609/aaai.v34i10.7244

Xu, Z., Li, J., Cheng, S., Rui, X., Zhao, Y., He, H., & Xu, L. (2024). Wildfire risk prediction: A review. *arXiv preprint arXiv:2405.01607*.

Zhai, J., Zhang, S., Chen, J., & He, Q. (2018). Autoencoder and its various variants. In *2018 ieee international conference on systems, man, and cybernetics (sc)* (pp. 415–419).

Zhong, C., Cheng, S., Kasoar, M., & Arcucci, R. (2023). Reduced-order digital twin and latent data assimilation for global wildfire prediction. *Natural Hazards and Earth System Sciences*, *23*(5), 1755–1768. https://doi.org/10.5194/nhess-23-1755-2023

Zhou, X., Mahalingam, S., & Weise, D. (2005). Modeling of marginal burning state of fire spread in live chaparral shrub fuel bed. *Combustion and Flame*, *143*(3), 183–198. https://doi.org/10.1016/j.combustflame.2005.05.013

Zhou, Y., Ruige, K., & Sibo, C. (2024). The codes for xai tools for wildfire prediction and explanation. https://doi.org/10.5281/zenodo.14286931

29935210, 2025, 2, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000409 by Worcester Polytechnic, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License