Pre-Ignition Forest Fire Risk Prediction Using Multi-Temporal Vegetation Indices and Machine Learning: A Case Study from California

Jayantrao Mohite[1*], Suryakant Sawant[2], Ankur Pandit[3], Dineshkumar Singh[1]

[1]Tata Consultancy Services, Mumbai, India - (jayant.mohite, dineshkumar.singh)@tcs.com
[2]Tata Consultancy Services, Pune, India - suryakant.sawant@tcs.com
[3]Tata Consultancy Services, Indore, India - ankur.pandit@tcs.com

Keywords: Forest fire prediction, Machine learning, Logistic Regression, Dynamic vegetation indices, California wildfires, Satellite data analysis.

Abstract

This study proposes a machine learning-based approach for predicting forest fire occurrences one month in advance in high-risk regions of California. Multiple algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), and a Hybrid LSTM + LR model, were systematically evaluated using dynamic vegetation indices such as Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Gross Primary Productivity (GPP), Leaf Area Index (LAI), along with their derivative components including trend and Exponential Moving Average (EMA). Static topographical and forest attributes such as slope, elevation, forest type, and forest group were also incorporated. Among the evaluated models, LR demonstrated the most consistent performance, achieving a validation Area Under the Curve (AUC) of 0.90 in Scenario F, which combined static variables with trend and EMA-based vegetation features. Experiments with varying historical windows (12, 9, and 6 months) showed that a 12-month period provided the best predictive accuracy by effectively capturing seasonal and long-term vegetation dynamics. The model was further validated using independent fire datasets from 2021 and 2024, achieving an AUC of 0.84 in 2021. However, performance declined to 0.64 in 2024, likely due to changes in satellite data products, spatial resolution, and forest phenology. The findings indicate that integrating long-term vegetation trends and EMA-based dynamic indices with static environmental variables can provide a robust and computationally efficient framework for forest fire prediction. However, future models must address data variability and environmental shifts to sustain long-term predictive performance.

1. Background and Introduction

Wildfire activity across the western United States has intensified in recent decades, emerging as a significant natural hazard characterized by increasing frequency, severity, and spatial expansion under the influence of both climate change and human activities. Factors such as prolonged droughts, rising temperatures, and declining humidity have heightened vegetation dryness and expanded the length of fire seasons (Abatzoglou and Williams, 2016, Jolly et al., 2015). California, characterized by its Mediterranean climate, steep terrain, and diverse forest ecosystems, remains one of the most fire-prone regions globally (Westerling, 2016, Bowman et al., 2009). These fires not only cause loss of biodiversity and property but also contribute significantly to atmospheric greenhouse gas emissions, altering regional carbon budgets (Running et al., 2004, Zhao et al., 2005).

Conventional fire-risk assessment approaches are largely based on meteorological variables or empirical fire-danger indices, including the National Fire Danger Rating System (Cohen and Deeming, 1985) and the Canadian Fire Weather Index (Van Wagner, 1987). Although these indices effectively describe daily or short-term ignition potential, they do not adequately capture vegetation dynamics, fuel condition, or phenological changes that precede ignition (Chuvieco et al., 2010, Bedia et al., 2015, Vilar et al., 2010). Physically based spread simulators such as FARSITE (Finney, 2004) are effective for forecasting post-ignition behavior but are not designed for regional-scale pre-ignition probability assessment. Hence, there is a growing need for interpretable, data-driven approaches that integrate multi-temporal vegetation information with static physiographic and ecological variables to anticipate potential fire occurrences before ignition.

Recent studies have demonstrated the value of remote-sensing-derived indices such as NDVI, NDWI, LAI, and GPP for quantifying vegetation greenness, moisture, and productivity, which directly influence fuel build-up and curing (Didan, 2015, Gao, 1996, Myneni et al., 2002, Myneni et al., 2015). Integrating such dynamic indicators with topographic and forest attributes through machine learning has improved pre-fire prediction accuracy (Rodrigues and de la Riva, 2014, Michael et al., 2021, Jing et al., 2024). However, several limitations persist such as inconsistent lead-time definitions, imbalance between fire and non-fire samples, the use of raw spectral bands without contextual features, and the challenge of overfitting in highly nonlinear models (Chen and Guestrin, 2016, Cortes and Vapnik, 1995). These limitations highlight the need for frameworks that preserve model interpretability, clearly define temporal lead times, and remain robust across spatial and temporal variability.

In this study, we propose a pre-ignition forest-fire risk prediction framework for California that explicitly models vegetation condition in the months preceding ignition. Static features (elevation, slope, aspect, forest type, and forest group) are combined with multi-temporal vegetation indices (NDVI, NDWI, LAI, and GPP) aggregated to monthly 1-km grids. For each index, we compute two derivative features namely trend and exponential moving average (EMA), within 12, 9, and 6-month lag windows to capture both the direction and recency of vegetation change. Four algorithms such as Logistic Regression (LR), Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), and a hybrid LSTM + LR are systematically evaluated across seven feature-set scenarios (A–G) that integrate static and dynamic predictors to identify the optimal configuration for operational early-warning use. Non-fire points were generated through spatial–temporal matching with a minimum 1-km exclusion buffer to ensure independence and balance between classes. The models were trained on data from 2010–2018, validated on 2019–2020, and tested on independent 2021 data. Additional evaluation on 2024 data using newer MODIS collections examines the sensitivity of the approach to sensor-product and phenological shifts.

The main objectives of this research are therefore to: 1. Develop an interpretable, lead-time explicit machine-learning framework for pre-ignition forest-fire prediction; 2. Compare the contribution of different vegetation indicators, lag-window lengths, and algorithmic families; and 3. Evaluate temporal generalization and understand the reasons for performance degradation under evolving data conditions.

This framework aims to support regional early warning and risk-assessment systems by providing a scalable, data-driven approach that integrates multi-source static and dynamic information while maintaining interpretability and operational practicality.

2. Materials and Methods

2.1 Study Area

The analysis was conducted over forested areas of California, USA, exhibiting strong climatic gradients, complex terrain, and high inter-annual variability in vegetation and moisture. The state's Mediterranean climate includes wet winters and dry summers which creates alternating phases of fuel accumulation and curing that predispose large areas to wildfire (Westerling, 2016, Bowman et al., 2009). These ecological and physiographic contrasts, together with the diverse fire regimes observed between 2010 and 2020, provide an ideal test region for developing and validating pre-ignition risk prediction models (Figure 1).

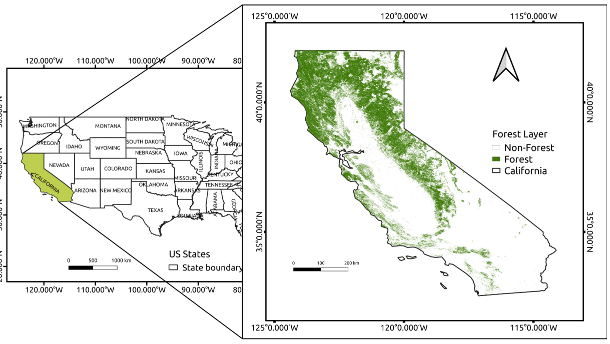

Figure 1. Study Area

2.2 Fire and Non-Fire Data

2.2.1 Forest fire events The forest fire locations used in this study were obtained from the United States Forest Service (USFS) Fire Occurrence Database, specifically from the Fire Occurrence Dataset (FOD), version 6.0 (Short, 2022). The FOD provides a comprehensive record of wildfire occurrences across the United States, including spatial data on fire ignition points along with relevant attributes such as fire size, cause, and geographic location. This dataset is widely used in wildfire research due to its consistency and coverage over multiple decades and provides a rich historical dataset for spatial and temporal fire analysis. Figure 2 shows the cumulative number of forest fire events across each state of the USA. California was chosen as the study area due to its unique combination of climatic, geographic, and ecological factors that contribute to a high number of fire events. The dataset was subsequently filtered to retain only fire occurrences within California during the period 2010–2020. As part of this analysis, we restricted our focus to fire events attributed to natural causes. Furthermore, we refined the dataset to include only seasonal fire events, typically from May to October of each year (Figure 3). This period represents the highest risk of wildfires due to climatic conditions conducive to fire spread.

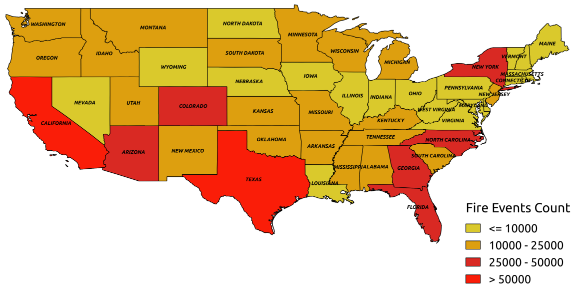

Figure 2. State-wise Cumulative Number of Fire Events in US States during 2010-2020

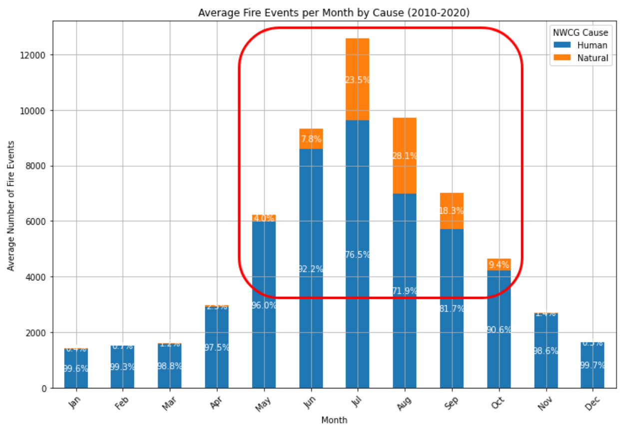

Figure 3. Fire Season in California

2.2.2 Creation of non-fire events Non-fire points were generated based on the spatial and temporal characteristics of the fire events to address the issue of imbalanced datasets. Using the fire event layer and the forest type layer, non-fire points were selected within forested regions to ensure that they represent non-fire conditions. These selected points were at least 1 km away from any recorded fire point to avoid overlap and ensure a clear distinction between fire and non-fire locations. Additionally, the non-fire points were temporally matched to the fire events to ensure similar time distribution. For example, if 20 fire events occurred in June 2012, an equal number of non-fire points were generated for the same period, with dates closely matching those of the fire events. This method ensured that the non-fire data had the same spatio-temporal characterist-

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-1403-2026 |

ics as the fire data, thus removing any potential bias that could arise from unbalanced temporal or spatial distribution, a challenge often noted in previous wildfire prediction studies (Yang et al., 2021). Balancing datasets has been shown to reduce the risk of overfitting to the majority class, particularly in scenarios where fire events significantly outnumber non-fire events (Chen et al., 2024).

2.2.3 Static Variables — Topographic and Forest Attributes Topographic and vegetation-type characteristics were incorporated as static predictors to capture spatial variability that persists through time. Elevation, slope, and aspect influence local temperature, radiation, and moisture regimes that modulate ignition potential and fire spread (Gallant and Wilson, 2000, Finney, 2004). Higher elevations and steep, south facing slopes generally exhibit faster fuel drying compared with shaded or gentle terrain. Two categorical layers namely forest type and forest group represent compositional and structural differences in vegetation, including canopy density, understory composition, and management history. All static variables were harmonized to a 1 km grid and joined to each fire and non-fire point, forming a consistent set of physiographic and ecological descriptors.

2.2.4 Dynamic Variables — Vegetation Indices Dynamic predictors describe the seasonal and inter-annual variability of vegetation condition and productivity that precede ignition. Four satellite-derived indices were selected based on proven relevance to fuel greenness and moisture dynamics.

Normalized Difference Vegetation Index (NDVI) quantifies canopy greenness and photosynthetic activity, serving as a proxy for vegetation vigor and live-fuel accumulation (Didan, 2015). Declines in NDVI preceding fire events often indicate stress or senescence.

Normalized Difference Water Index (NDWI) highlights vegetation water content and surface moisture (Gao, 1996, Didan, 2015). Reduced NDWI values correspond to fuel drying and increased flammability.

Leaf Area Index (LAI) represents one-sided leaf area per unit ground area, capturing canopy density and potential fuel continuity (Myneni et al., 2002, Myneni et al., 2015).

Gross Primary Productivity (GPP) measures carbon uptake and overall vegetation productivity, integrating physiological responses to light, temperature, and water availability. Lower GPP can signal stress and reduced moisture, precursors to ignition (Running et al., 2004, Zhao et al., 2005).

Monthly composites for all indices were derived from MODIS Terra–Aqua products and aggregated to 1 km resolution for each sample location.

2.2.5 Preparation of Vegetation Time Series and Engineered Features For every index (NDVI, NDWI, LAI, GPP), time-series were constructed for the preceding 12, 9, and 6-month periods relative to each event date. Two derived features were computed per index and lag window: (i) trend, representing the direction and magnitude of change through time, and (ii) exponential moving average (EMA), which emphasizes recent conditions while retaining continuity. These compact metrics encapsulate seasonal phenology and moisture dynamics essential for pre-ignition modeling while reducing dimensionality and noise (Michael et al., 2021).

2.2.6 Machine-Learning Algorithms Four supervised learning algorithms were employed to represent different modeling paradigms and complexity levels. Logistic Regression (LR) provides a statistically interpretable baseline that expresses the probability of fire occurrence as a logistic function of predictor variables (Cox, 1958, Rodrigues and de la Riva, 2014). Regularization was applied to control multicollinearity and prevent overfitting. Support Vector Machine (SVM) constructs a maximum margin decision boundary in a transformed feature space and can model nonlinear relationships between predictors and fire probability (Cortes and Vapnik, 1995). Extreme Gradient Boosting (XGB), a boosted-tree ensemble, captures higher-order interactions efficiently while incorporating shrinkage and subsampling for regularization (Chen and Guestrin, 2016). A hybrid Long Short-Term Memory (LSTM) plus Logistic Regression model was also implemented to represent temporal dependencies explicitly: an LSTM encoder summarizes monthly vegetation sequences into latent features, which are concatenated with static predictors and passed to an LR classifier. This configuration preserves temporal context while maintaining interpretability at the final stage. All models were trained with hyper-parameter tuning via grid search and five-fold cross-validation.

2.2.7 Overall methodological approach In this study, we developed a comprehensive modeling framework (Figure 4) to predict forest fire risk in the California region. The dataset consisted of 11,123 points, with 5,791 non-fire and 5,332 fire events, spanning the period from 2010 to 2020. The features used for modeling included both static and dynamic variables. Static features comprised topographical attributes such as elevation, slope, and aspect, along with forest attributes including forest type and forest group. Moreover, dynamic vegetation indices such as NDVI, NDWI, GPP, and LAI - were also included to capture vegetation health and moisture variations over time. These dynamic indices were extracted for the twelve months preceding each event, and monthly means were calculated. Additionally, trends and exponential moving averages (EMA) were estimated for each index, providing insights into both long-term and short-term vegetation dynamics. To evaluate the impact of these features on fire occurrence, multiple feature scenarios were examined by combining static variables with various configurations of dynamic indices, their associated trends, and EMA (Table 1). This was done to get a clear understanding of how both static and dynamic variables contribute to fire risk. We implemented four machine learning algorithms to model the probability of fire occurrence: Logistic Regression (LR), Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), and a hybrid model combining Long Short-Term Memory (LSTM) networks with Logistic Regression (LSTM + LR). Model development was carried out using the data from 2010 to 2018, with hyperparameter tuning performed using three-fold cross-validation. The models were subsequently validated on data from 2019 to 2020. Performance was evaluated using standard classification metrics, including the AUC, and ROC-AUC, to ensure robustness and accuracy. This systematic evaluation of feature scenarios, historical time windows and algorithms can help to identify the best-performing model for predicting forest fire occurrences one month in advance, based on the integration of static landscape features and dynamic vegetation conditions, thereby providing a valuable tool for fire risk assessment.

2.2.8 Validation using Fire data of 2021 from US-NIFC To further evaluate the model's generalization ability and robustness, independent validation was performed using fire data

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-1403-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 1. Various feature set scenarios used for modeling

SN	Scenario	Feature set
1	A	Static features (Topography, Forest))
2	B	A + df(NDVI, NDWI)
3	C	B + TS(NDVI, NDWI)
4	D	A + df(NDVI, NDWI, GPP)
5	E	D + TS(NDVI, NDWI, GPP)
6	F	A + df(NDVI, NDWI, GPP, LAI)
7	G	F + TS(NDVI, NDWI, GPP, LAI)

Note: df – derivative features such as trend and exponential moving average (EMA); TS – monthly time-series indices.

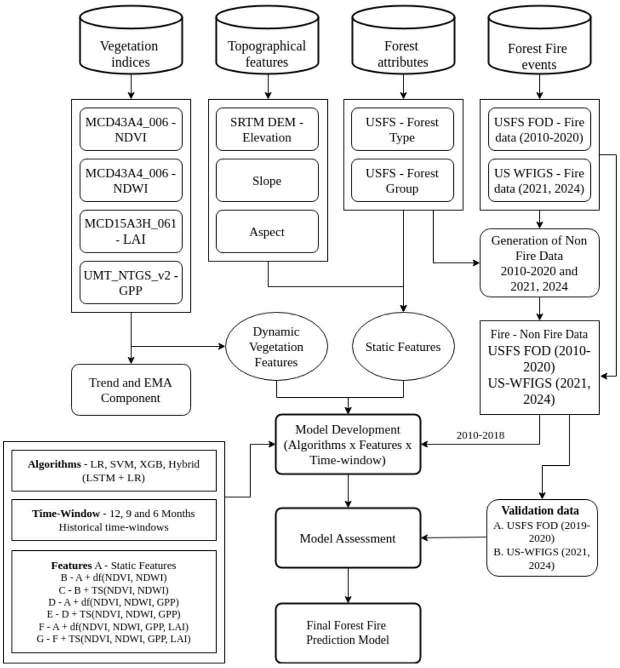

Figure 4. Forest fire prediction modeling framework

from the United States National Interagency Fire Center (US-NIFC) for the year 2021 and 2024 (National Interagency Fire Center (NIFC), 2026). This dataset contains recorded fire events across the United States, was filtered to focus on the fire season from May to October 2021 (Figure 5) and May to September 2024. As the Fire Program Analysis (FPA) data used for model training did not cover the period after 2020, this independent dataset provided an opportunity to assess the model’s performance on completely unseen data from a different source. Using a similar approach as previously employed for generating non-fire events, non-fire points were created by ensuring spatial separation from fire events and maintaining comparable temporal distribution. Static variables, including topographical and forest attributes, were estimated for each point, while dynamic vegetation indices (NDVI, NDWI, GPP, and LAI) were extracted for the twelve months preceding each event. Trends and exponential moving averages (EMA) were also calculated for these dynamic indices, following the same methodology used during the initial model development. This independent validation allowed for a robust assessment of the model’s performance when applied to new data.

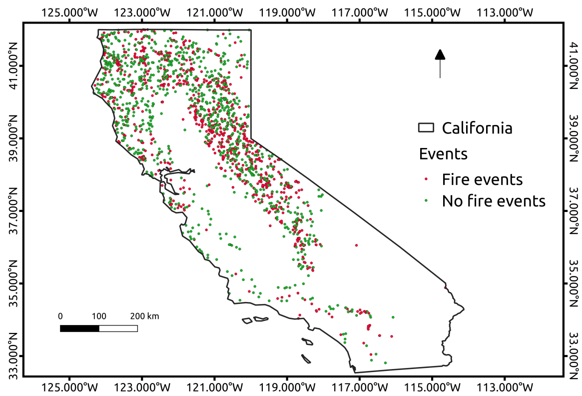

Figure 5. Fire and no fire events from 2021

3. Results

3.1 Exploratory Assessment of Variables

An initial analysis of static and dynamic predictors was carried out to understand their separability between fire (FF) and non-fire (NF) samples. Topographic variables showed moderate influence. Fire points were more frequently observed on steeper slopes and at higher elevations, indicating terrain’s role in fuel drying and airflow dynamics. The aspect variable revealed subtle directional preferences, with a slight dominance of south and west facing slopes—consistent with higher solar exposure and faster desiccation (Finney, 2004). Figure 6 illustrates the boxplot distribution of elevation, slope, and aspect for fire and non-fire classes.

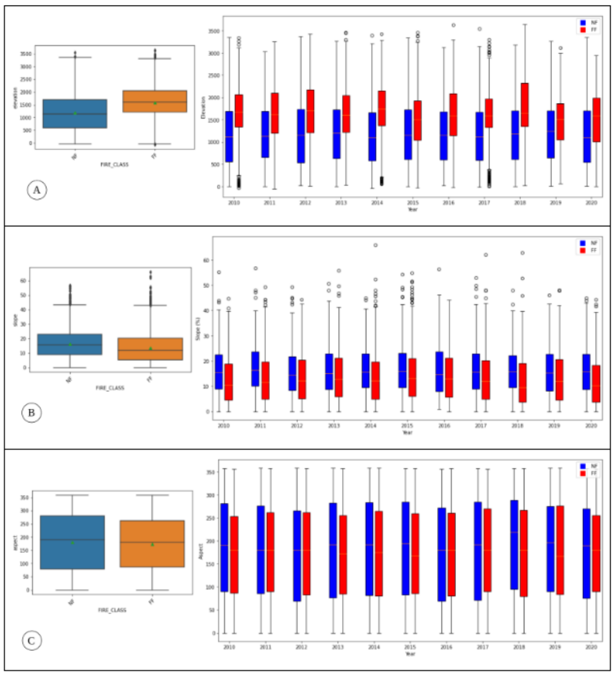

Figure 6. Yearly and combined box plot for static variables

Vegetation indices exhibited clear temporal contrasts preceding fire events. NDVI and NDWI values typically declined during the final three to five months before ignition, reflecting vegetation stress and moisture loss, whereas LAI and GPP showed seasonal peaks earlier in the year followed by a drop toward the dry months. This behavior highlights the importance of both greenness and water-related dynamics as pre-ignition indicators (Didan, 2015, Gao, 1996, Running et al., 2004). Figure 7 presents monthly trajectories of these indices for representative FF and NF samples.

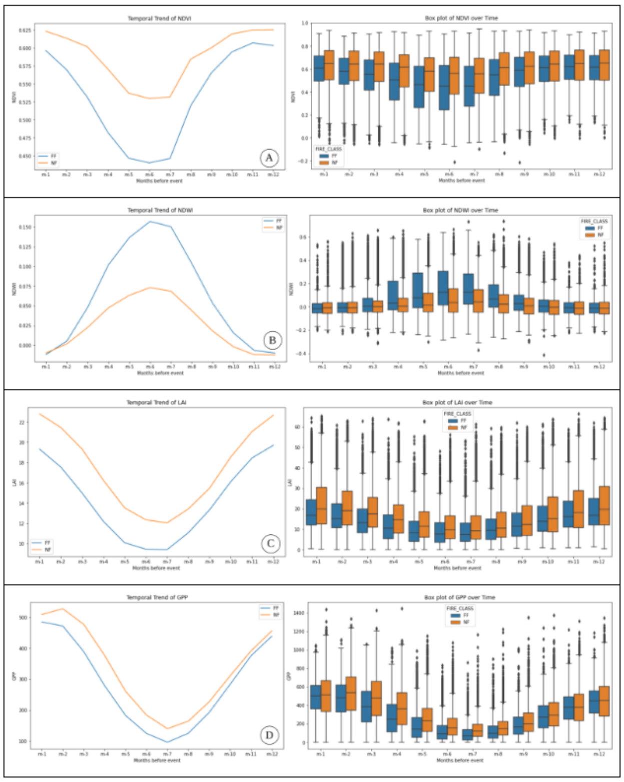

Figure 7. Yearly and combined box plot for dynamic variables

3.2 Model Performance across Algorithms and Scenarios

Model performance was evaluated for all four algorithms (LR, SVM, XGB, LSTM + LR) across the seven feature-set scenarios (A–G) shown in Table 1. The comparison highlights how the inclusion of dynamic vegetation information improves predictive capability over static variables alone.

3.2.1 Algorithm comparison Across all scenarios, Logistic Regression (LR) consistently provided the most stable and generalizable results. Across all feature scenarios, LR shows consistent results with minimal drop in AUC between the test and validation sets (Table 2). For example, in Scenario F (twelve months), LR achieves a test AUC of 0.84 and a validation AUC of 0.90, which is one of the highest validation AUCs across all algorithms and feature scenarios. This suggests that LR exhibits strong robustness and generalizes effectively to unseen data, thereby reducing the risk of overfitting, which is crucial for reliable forest fire prediction. On the contrary, Extreme Gradient Boosting (XGB), while showing good performance on the test set (e.g., 0.99 in Scenario F), shows a significant drop in AUC on the validation set (0.68 in the same scenario), suggesting possible overfitting. XGB captures the detailed patterns in the training data well but fails to generalize these patterns to unseen data, making it less reliable for predictions on new forest fire data. Support Vector Machine (SVM), though relatively stable, generally shows lower AUC values compared to LR. For instance, in Scenario F (twelve months), SVM achieves a test AUC of 0.82 and a validation AUC of 0.83. While SVM performs better than XGB in terms of generalization, it still lags slightly behind LR, making it a less optimal choice. Finally, the Hybrid LSTM + LR model shows that while it performs well (e.g., 0.85 validation AUC in Scenario F with twelve months of data), the LSTM component does not add substantial value compared to LR alone. Its performance is comparable to that of LR, indicating that the additional complexity introduced by the hybrid model offers only marginal performance improvements.

3.2.2 Sensitivity to Historical Window Length The evaluation across different historical data windows (twelve, nine, and six months), reveals that twelve months of data consistently provides the best AUC results across algorithms and feature scenarios (Table 2). In Scenario F, for example, LR achieves a validation AUC of 0.90 with twelve months of data. This window allows the model to capture critical seasonal trends in vegetation indices, which are essential for accurate fire prediction. The nine-month period also provides competitive results, though slightly lower than the twelve-month period. For example, LR’s validation AUC in Scenario F is 0.87 with nine-months of data. This suggests that while a nine-month window is sufficient to capture important trends, the additional information provided by the twelve-month window gives the model a better understanding of long-term changes in the forest. The 6-month window results in a noticeable drop in AUC values across all models. For example, in Scenario F, LR’s validation AUC drops to 0.81, indicating that a shorter time window does not capture enough seasonal variability or long-term trends in vegetation indices, which are crucial for accurate forest fire prediction. This further reinforces the choice of using twelve months of data as the optimal historical window for prediction.

3.2.3 Feature Importance Analysis Among the various feature scenarios, Scenario F (which includes static features and derivative features such as trend and EMA of NDVI, NDWI, GPP, and LAI) consistently provides the best balance between validation AUC and computational complexity (Table 2). In Scenario F with twelve months of data, LR achieves a validation AUC of 0.90, making it the best-performing feature scenario (Table 2, row highlighted in brown background). Scenario G, which adds time-series features of the indices on top of the derivative features from Scenario F, slightly improves the test AUC but with a cost of significant computational overhead. For example, in Scenario G with twelve months of data, LR achieves a validation AUC of 0.90, the same as Scenario F, but with additional computational requirements due to the inclusion of full time-series data. Given the minimal performance improvement, the computational cost of Scenario G may not be justified for practical applications. Therefore, Scenario F is the optimal feature selection scenario as it captures the necessary trends and patterns in vegetation indices without the added computational expense of processing full time-series data, making it an efficient choice for forest fire prediction.

The feature-importance analysis for Scenario F (Logistic Regression model) revealed that the exponential moving averages (EMA) of NDVI and NDWI were the most influential predictors of forest fire occurrence (Figure 8). These indicators effectively captured long-term vegetation stress and moisture depletion patterns, which are critical precursors to ignition. Specifically, NDVI_ema showed the highest importance score, followed by NDWI_ema, confirming that persistent declines in greenness

Table 2. Performance of various models across historical windows

SC	Features	Window (months)	LR		SVM		XGB		Hybrid	
			Test	Val	Test	Val	Test	Val	Test	Val
A	Static	12	0.69	0.70	0.65	0.65	0.65	0.62		
		9	0.69	0.70	0.65	0.65	0.65	0.62		
		6	0.69	0.70	0.65	0.65	0.65	0.62		
B	A + df(NDVI + NDWI)	12	0.79	0.78	0.76	0.77	0.95	0.68		
		9	0.78	0.77	0.74	0.72	0.96	0.69		
		6	0.75	0.76	0.70	0.67	0.96	0.64		
C	B + TS(NDVI, NDWI)	12	0.80	0.80	0.76	0.74	0.99	0.76	0.78	0.76
		9	0.80	0.79	0.77	0.75	0.99	0.77	0.77	0.76
		6	0.77	0.78	0.74	0.74	0.99	0.72	0.74	0.76
D	A + df(NDVI, NDWI, GPP)	12	0.82	0.83	0.81	0.81	0.99	0.65		
		9	0.82	0.83	0.77	0.77	0.99	0.65		
		6	0.78	0.77	0.75	0.73	0.99	0.61		
E	D + TS(NDVI, NDWI, GPP)	12	0.84	0.85	0.81	0.81	0.96	0.73	0.80	0.82
		9	0.83	0.84	0.80	0.79	0.96	0.71	0.78	0.80
		6	0.79	0.79	0.76	0.73	0.96	0.65	0.76	0.76
F	A + df(NDVI, NDWI, GPP, LAI)	12	0.84	0.90	0.82	0.83	0.99	0.68		
		9	0.83	0.87	0.82	0.79	0.99	0.64		
		6	0.79	0.81	0.78	0.74	0.99	0.60		
G	F + TS(NDVI, NDWI, GPP, LAI)	12	0.86	0.90	0.84	0.85	0.96	0.74	0.84	0.85
		9	0.84	0.88	0.82	0.81	0.96	0.72	0.82	0.83
		6	0.81	0.82	0.79	0.77	0.96	0.63	0.77	0.79

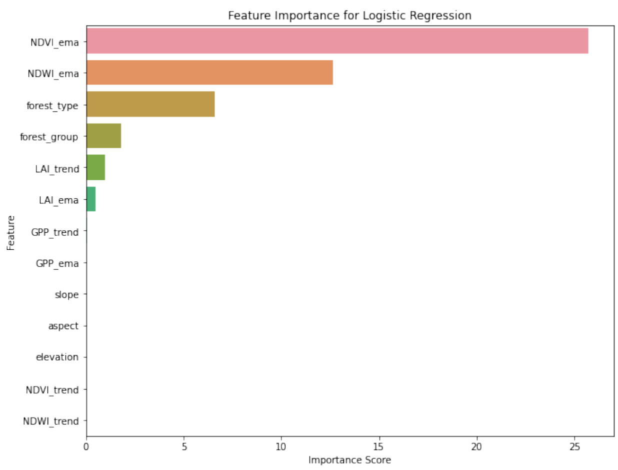

Figure 8. Importance of various features from best LR model

and water content make forests more susceptible to fire. The forest type variable also contributed meaningfully, indicating that vegetation composition influences ignition potential, while forest group provided broader contextual information on ecosystem characteristics.

Other vegetation indices such as LAI and GPP, though less dominant, added complementary insights into canopy structure and ecosystem productivity. LAI trend and LAI ema reflected thinning canopy conditions, whereas GPP-based features captured physiological stress responses linked to reduced moisture. In contrast, static features like slope, elevation, and aspect exhibited negligible influence, suggesting that dynamic vegetation conditions outweigh topographic effects in determining pre-fire risk. The negligible contribution of NDVI trend and NDWI trend further indicates that their signals are already encapsulated within the EMA metrics, which smooth temporal variations and highlight meaningful long-term drying trends.

3.2.4 Independent Temporal Validation To assess the generalization ability of final model, an independent validation was performed using fire data from the US-WFIGS dataset for the years 2021 and 2024. This allowed the model to be tested on completely unseen data, distinct from the 2010–2020 period used for training and internal validation. The results (Figure 9) indicate that the model achieves an AUC of 0.84 on the 2021 data, demonstrating strong predictive performance and confirming the model's ability to generalize to new fire events beyond the training period. However, when applied to the 2024 data, the model's performance shows a noticeable decline, with the AUC dropping to 0.64 (Figure 9). Several factors may explain this performance drop in 2024: A. Changes in MODIS Products: Between 2021 and 2024, several MODIS products used to derive key indices such as NDVI, NDWI, and GPP were either discontinued or replaced with alternative products that have different compositing mechanisms, spatial and temporal resolution. Such changes could lead to inconsistencies in the derived features and likely contributed to the increased misclassification of fire and non-fire points in 2024, as the model was trained on data with different characteristics. B. Spatial Resolution Changes: The spatial resolution of some of the indices changed between the original MODIS products and their replacements. This could have introduced discrepancies in how fine-scale forest dynamics were captured. The model, having been trained on a specific spatial resolution (e.g., 1km), might have struggled to adapt to new resolutions, affecting its ability to capture small-scale variations that are crucial for predicting fire events. C. Phenological and Environmental Changes:

Another contributing factor could be the shifts in forest phenology due to environmental changes, such as increased drought, changes in vegetation cycles, or alterations in fire-prone areas. Over time, forests undergo changes in biomass, leaf area, and moisture content, which directly affect their fire susceptibility. These evolving conditions may have caused a shift in the feature space, making the forest fire dynamics in 2024 different from those in previous years, thus impacting the model's accuracy.

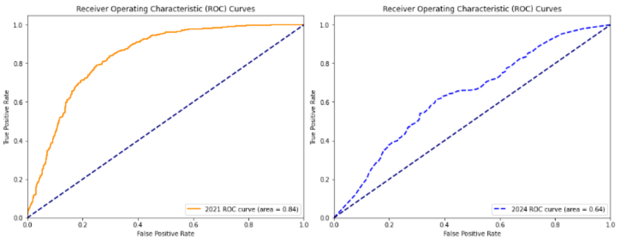

Figure 9. ROC AUC Curve for independent data from 2021 and 2024

4. Discussion and Conclusions

The results demonstrate that combining multi-temporal vegetation dynamics with static topographic and forest attributes provides a robust basis for pre-ignition forest-fire risk prediction. Among the four algorithms evaluated, Logistic Regression (LR) consistently achieved the best performance and generalization capability, with AUC values around 0.90 on internal validation and 0.84 on independent 2021 testing. Its stability across years confirms that a regularized linear formulation is sufficient to capture the largely monotonic relationship between vegetation condition and ignition likelihood, while avoiding the overfitting tendencies observed in SVM, XGB, and LSTM + LR models (Rodrigues and de la Riva, 2014, Chen and Guestrin, 2016, Cortes and Vapnik, 1995).

The superiority of the 12-month window emphasizes the need to encompass a full phenological cycle of vegetation build-up, senescence, and drying before ignition. Shorter 6 or 9-month windows might omit early-season growth phases that govern total fuel availability, thereby reducing predictive strength. This finding is consistent with prior studies linking annual vegetation cycles to fire probability in Mediterranean ecosystems (Jolly et al., 2015, Michael et al., 2021).

Dynamic vegetation indicators such as NDWI and NDVI emerged as the most influential predictors. Both indices capture greenness and moisture variability that translate directly to live-fuel moisture content, a critical determinant of ignition susceptibility (Gao, 1996, Didan, 2015, Yebra et al., 2013). Among them, NDWI trend and NDWI EMA ranked highest, confirming that recent drying and cumulative moisture loss are more decisive than absolute vegetation greenness. Additional indices such as LAI and GPP contributed secondary information related to canopy structure and productivity, providing complementary evidence of vegetation stress (Myneni et al., 2002, Running et al., 2004, Zhao et al., 2005). While forest type and forest group retained measurable influence, topographic variables (slope, aspect, elevation) were relatively less discriminative once vegetation condition was accounted for.

The sharp performance decline on 2024 data (AUC 0.64) illustrates the impact of sensor-product and environmental shifts on predictive stability. Differences between MODIS Collections 006 and 061, resolution changes, and adjustments in atmospheric corrections can alter vegetation-index magnitudes and temporal consistency. Moreover, evolving phenological patterns under climate variability may shift feature distributions beyond the model's trained domain (Thorne et al., 2018). These findings underscore the importance of implementing periodic recalibration and domain-adaptation mechanisms to ensure consistent model performance across sensors and years.

From an operational perspective, the study shows that interpretable, computationally light models can effectively deliver early-warning capability when supported by well-curated temporal features. The use of trend and exponential moving average (EMA) condenses lengthy time-series into features that preserve ecological meaning and enable rapid inference. Such models are suitable for large-area monitoring, allowing fire-management agencies to identify zones of elevated ignition potential linked to vegetation stress or anomalous drying. Because predictors are physically interpretable, the framework also supports diagnostic insights rather than black-box outputs which is also beneficial for real-world decision support. In conclusion, this research establishes an efficient and transparent method for pre-ignition forest-fire risk prediction using satellite-derived vegetation dynamics and static physiography. Key outcomes include: 1. **Model performance**: Logistic Regression provided the highest and most stable predictive accuracy. 2. **Temporal context**: A 12-month history of vegetation dynamics was optimal for capturing pre-fire evolution. 3. **Feature relevance**: Moisture-related indicators, particularly NDWI-based trend and EMA, dominated predictive importance. 4. **Transferability**: The model generalized well to 2021 data but declined for 2024 due to product and phenological shifts, indicating the need for retraining and harmonization.

The study highlights that, with consistent preprocessing and periodic calibration, the proposed framework can be scaled for regional early-warning and integrated with meteorological, soil-moisture, or SAR-based inputs to further enhance reliability. Future work will focus on sensor harmonization, cross-sensor fusion, and adaptive retraining to maintain performance under changing environmental and data-collection conditions.

References

Abatzoglou, J. T., Williams, A. P., 2016. Impact of anthropogenic climate change on wildfire across western US forests. *Proceedings of the National Academy of Sciences*, 113(42), 11770–11775.

Bedia, J., Herrera, S., Camia, A., Moreno, J. M., Gutiérrez, J. M., 2015. Forest fire danger projections in the Mediterranean using ENSEMBLES regional climate change scenarios. *Climatic Change*, 122(1-2), 185–199.

Bowman, D. M. J. S., Balch, J. K., Artaxo, P., Bond, W. J., Carlson, J. M., Cochrane, M. A. et al., 2009. Fire in the Earth System. *Science*, 324(5926), 481–484.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, X., Tian, Y., Zheng, C., Liu, X., 2024. AutoST-Net: A spatiotemporal feature-driven approach for accurate forest fire spread prediction from remote sensing data. *Forests*, 15(4), 705.

Chuvieco, E., Aguado, I., Yebra, M., Nieto, H., Salas, J., Martín, M. P., Vilar, L., Martínez, J., Martín, S., Ibarra, P., 2010. Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. *Ecological Modelling*, 221(1), 46–58.

Cohen, J. D., Deeming, J. E., 1985. The national fire-danger rating system: Basic equations. *USDA Forest Service General Technical Report PSW-82*.

Cortes, C., Vapnik, V., 1995. Support-vector networks. *Machine Learning*, 20(3), 273–297.

Cox, D. R., 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B*, 20(2), 215–242.

Didan, K., 2015. MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006. *NASA EOSDIS Land Processes DAAC*.

Finney, M. A., 2004. FARSITE: Fire Area Simulator—Model development and evaluation. *USDA Forest Service Research Paper RMRS-RP-4*, 1–47.

Gallant, J. C., Wilson, J. P., 2000. Primary topographic attributes. *Terrain Analysis: Principles and Applications*, John Wiley & Sons, 51–85.

Gao, B.-C., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3), 257–266.

Jing, X., Li, X., Zhang, D., Liu, W., Zhang, W., Zhang, Z., 2024. Forecast zoning of forest fire occurrence: A case study in southern China. *Forests*, 15(2), 265.

Jolly, W. M., Cochrane, M. A., Freeborn, P. H., Holden, Z. A., Brown, T. J., Williamson, G. J., Bowman, D. M. J. S., 2015. Climate-induced variations in global wildfire danger from 1979 to 2013. *Nature Communications*, 6, 7537.

Michael, Y., Helman, D., Glickman, O., Gabay, D., Brenner, S., Lensky, I. M., 2021. Forecasting fire risk with machine learning and dynamic information derived from satellite vegetation index time-series. *Science of The Total Environment*, 764, 142844.

Myneni, R. B., Hoffman, S., Knyazikhin, Y., Privette, J. L., Glassy, J., Tian, Y. et al., 2002. Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sensing of Environment*, 83(1-2), 214–231.

Myneni, R. B., Knyazikhin, Y., Park, T., 2015. MOD15A2H MODIS/Terra Leaf Area Index/FPAR 8-Day L4 Global 500m SIN Grid V006. *NASA EOSDIS Land Processes DAAC*.

National Interagency Fire Center (NIFC), 2026. Wildland Fire Incident Locations. https://data-nifc.opendata.arcgis.com/datasets/nifc::wildland-fire-incident-locations/explore?location=0.000000Accessed: 2026-05-14.

Rodrigues, M., de la Riva, J., 2014. An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software*, 57, 192–201.

Running, S. W., Nemani, R. R., Heinsch, F. A., Zhao, M., Reeves, M., Hashimoto, H., 2004. A continuous satellite-derived measure of global terrestrial primary production. *BioScience*, 54(6), 547–560.

Short, K. C., 2022. Spatial wildfire occurrence data for the United States, 1992-2020 [FPA_FOD_20221014].

Thorne, P. W., Allan, R. P., Bojinski, S., Easterbrook, S., Hegerl, G. C. et al., 2018. Toward a global land and ocean surface climate observing system. *Bulletin of the American Meteorological Society*, 99(9), 1853–1874.

Van Wagner, C. E., 1987. Development and structure of the Canadian Forest Fire Weather Index System. *Canadian Forestry Service, Forestry Technical Report*, 35, 1–37.

Vilar, L., Woolford, D. G., Martell, D. L., Martín, M. P., Chuvieco, E., 2010. A comparison of remote sensing and GIS-based forest fire danger models in Mediterranean environments. *International Journal of Wildland Fire*, 19(6), 758–773.

Westerling, A. L., 2016. Increasing western US forest wildfire activity: sensitivity to changes in the timing of spring. *Philosophical Transactions of the Royal Society B*, 371(1696), 20150178.

Yang, S., Lupascu, M., Meel, K. S., 2021. Predicting forest fire using remote sensing data and machine learning. *Proceedings of the AAAI conference on artificial intelligence*, 35number 17, 14983–14990.

Yebra, M., Dennison, P. E., Chuvieco, E., Riaño, D., Zylstra, P., Hunt, E. R., Danson, F. M., Qi, J., Jurdao, S., 2013. A global review of remote sensing of live fuel moisture content for fire danger assessment: moving towards operational products. *Remote Sensing of Environment*, 136, 455–468.

Zhao, M., Heinsch, F. A., Nemani, R. R., Running, S. W., 2005. Improvements of the MODIS terrestrial gross and net primary production global data set. *Remote Sensing of Environment*, 95(2), 164–176.