A Nationwide Benchmark for Wildfire Initial Attack Failure Prediction with Public Environmental Data

Runyang Xu[1]
Florida State University
rx23@fsu.edu

Xueqi Cheng[1]
Florida State University
xc25@fsu.edu

Yushun Dong*
Florida State University
yushun.dong@fsu.edu

Abstract

Initial attack (IA) is the first wildfire suppression phase, when agencies must quickly decide which fires may escape early control. Existing IA failure prediction studies often use non-public response records or regional settings, so it remains unclear how well public data available at fire discovery time can support IA failure prediction at national scale. We present WILDFIREIA, the first U.S. national-scale benchmark for IA failure prediction from environmental and contextual data available at fire discovery time. WILDFIREIA aligns 38,128 naturally caused FPA-FOD wildfire events with FIRMS/VIIRS thermal detections, gridMET weather and fire-danger variables, LANDFIRE vegetation, fuel, and topography, OpenStreetMap access features, and WorldPop population density. To prevent data leakage, the benchmark fixes the event unit, size-based label rule, chronological split, metrics, and forbidden-feature list, and excludes final fire size, containment timestamps, and post-discovery satellite detections from model inputs. We evaluate 16 representative models across tabular, temporal, spatial, and spatiotemporal families under the same protocol. Results show that public discovery-time data provides useful but incomplete signal for IA failure prediction: XGBoost achieves the best AUPRC of 53.3%; FIRMS/VIIRS is the least redundant source; and fuel is the strongest static predictor when dynamic observations are unavailable. We release preprocessing outputs and model-ready caches to support reproducible research on early wildfire risk assessment: https://github.com/LabRAI/WildfireIA#.

1 Introduction

The initial attack (IA) phase is the first suppression phase after a wildfire is discovered, when first-arriving firefighting resources attempt to halt or slow fire spread [54, 50]. At this stage, fire management agencies face an event-level triage problem: they must quickly predict whether a reported wildfire can be contained by the initial response or is likely to escape early control [50, 17, 85, 47]. If this risk is underestimated, a wildfire that still appears manageable may be treated just as a routine event, causing agencies to miss the narrow window for early escalation and forcing a shift from rapid containment to a larger, longer, and more costly suppression effort with greater operational complexity and life-safety risk [47, 64, 17, 6]. These stakes are becoming increasingly severe in U.S. recently as fire-weather seasons lengthen, large-fire activity increases, fuels become drier, and ignition patterns expand the conditions under which damaging wildfires can occur [82, 46, 28, 2, 10].

Previous IA failure prediction studies show that early suppression outcomes are associated with fire weather, fuels, terrain, access, location, and response-related factors [6, 54, 50, 17]. However, these studies leave two immediate barriers to building reliable early-warning models. First, many rely on operational variables such as dispatch time, crew size, resource availability, aircraft use, and deployment logs, which are important for explaining suppression outcomes but are often unavailable

*Corresponding author

Preprint.

in public U.S. national databases or unknown at fire discovery time [54, 47, 64, 50]. This makes their results difficult to reproduce and can mix the intrinsic early risk of a fire with information about the later suppression response. Second, existing evaluations are often regional or agency-specific, so their labels, response standards, source availability, and evaluation protocols are not directly comparable across studies [50, 15, 85]. As a result, agencies and researchers still lack clear evidence on how well public discovery-time data can predict IA failure, which sources add non-redundant signal, and which models are reliable under the same evaluation setting, limiting evidence-based early warning and resource prioritization. This uncertainty has practical consequences as wildfire activity continues to strain communities and response systems, with U.S. fires in 2026 already exceeding the 10-year average for the same year-to-date (YTD) period in both fire counts and burned area, and recent evidence showing that wildfire smoke is imposing growing public-health burdens [60, 27]. Therefore, a U.S. national public benchmark is urgently needed to turn IA failure prediction into a reproducible, leakage-controlled task where data sources and model families are compared under the same rules.

To tackle the research gaps, in this paper we present WILDFIREIA, the first U.S. national-scale benchmark for wildfire IA failure prediction from public environmental data available at fire discovery time. To address the dependence on non-public operational records, WILDFIREIA builds on 38,128 naturally caused FPA-FOD wildfire events as a national event backbone and aligns them with public discovery-time or static signals from FIRMS/VIIRS, gridMET, LANDFIRE, OpenStreetMap, and WorldPop [74, 58, 1, 53, 41, 78]. To address the lack of comparable evaluation protocols, WILDFIREIA fixes the event unit, size-based label rule, chronological split, forbidden-feature list, metrics, and model-ready representations, while excluding outcome-derived information such as final fire size, containment timestamps, and post-discovery satellite detections from all model inputs [74, 85, 50, 12, 79]. We focus on naturally caused wildfires because their early spread is more directly governed by ignition environment, fire weather, fuel, and terrain, whereas human-caused fire outcomes are more strongly shaped by human activity distributions and reporting patterns [10, 31, 32, 46]. Using this benchmark, we evaluate 16 representative models that cover both standard classical baselines and recent state-of-the-art architectures across tabular, temporal, spatial, and spatiotemporal prediction families [21, 23, 71, 77, 34]. The evaluation shows three key insights: (i) public discovery-time data contains meaningful but incomplete signal for IA failure prediction, with structured event-level models providing strong baselines; (ii) discovery-day FIRMS/VIIRS evidence is the least redundant input source, while fuel provides the strongest static fallback signal when dynamic observations are unavailable; and (iii) containment duration is only weakly explained by discovery-time inputs, indicating that post-discovery suppression actions and fire evolution remain dominant factors. Overall, our contributions are summarized as follows:

- **The first U.S. national benchmark for wildfire IA failure prediction.** We introduce WILDFIREIA, which standardizes IA failure prediction as an event-level task at fire discovery time, with fixed event units, labels, chronological splits, metrics, and leakage-control rules.
- **A leakage-controlled multi-source data infrastructure.** We align 38,128 naturally caused FPA-FOD wildfire events with public discovery-time and static signals from FIRMS/VIIRS, gridMET, LANDFIRE, OpenStreetMap, and WorldPop, and provide model-ready tabular, temporal, spatial, and spatiotemporal representations.
- **A broad baseline suite across prediction paradigms.** We evaluate 16 representative models covering classical baselines and recent architectures for tabular, temporal, spatial, and spatiotemporal learning, establishing a reproducible performance reference for future IA prediction research.
- **Benchmark-driven insights into early wildfire risk.** Our experiments show that public discovery-time data provides meaningful but incomplete signal for IA failure prediction, identify FIRMS/VIIRS as the least redundant discovery-day source and fuel as the strongest static fallback source, and show that containment duration is weakly explained by discovery-time inputs.

2 Benchmark Design and Data Construction

This section first defines the event-level IA failure prediction task and research questions, then describes the public data sources used in WILDFIREIA, and finally presents the canonicalization pipeline that converts these sources into leakage-controlled, model-ready artifacts.

Table 1: Dataset-source inventory for WILDFIREIA. Each row corresponds to a downloaded public data product or raster layer. "Canonical size" reports the scale after preprocessing into event-level, daily, or patch-level benchmark data. Area is reported in km^2 for the nominal source coverage.

Name	Source	Country	Area (km^2)	Task	Period	Spatial resolution	Update frequency
FPA-FOD	fire reports	USA	9,834,000	labels and metadata	1992–2020	event point	periodic release
FIRMS/VIIRS	fire signal	Worldwide	149,000,000	discovery-day thermal signal	2012–2026	375 m	near-real-time
gridMET	weather	USA	9,834,000	weather and fire danger	1979–2020	$\sim$4 km	daily
Landfire	fuel	USA	9,834,000	Fuel Disturbance, Vegetation, Canopy	2001–2024	30 m	versioned release
Landfire	vegetation	USA	9,834,000	Existing Vegetation	2001–2024	30 m	versioned release
Landfire	topography	USA	9,834,000	Elevation, Aspect, Slope Degrees	2001–2024	30 m	versioned release
OpenStreetMap	access	Worldwide	149,000,000	roads density and fire stations	2004–2026	vector	continuously updated
WorldPop	population	Worldwide	149,000,000	population density	2000–2026	$\sim$100 m	annual

2.1 Event-Level Task Design and Research Questions

This subsection defines the event unit, input rule, label construction, and research questions for WILDFIREIA. Let $i \in \{1, \ldots, n\}$ index wildfire events, and let t_i^d denote the reported discovery time of event i. Let a_i denote its final burned area, which is used only for label construction and never used as a model input. Let $\mathcal{C}$ denote the set of public source groups and $\mathcal{S} \subseteq \mathcal{C}$ denote any source subset. For event i, let $\boldsymbol{x}_i(\mathcal{S})$ denote the model-ready input constructed from sources in $\mathcal{S}$ using only public information available at or before t_i^d. The input $\boldsymbol{x}_i(\mathcal{S})$ may be tabular, temporal, spatial, or spatiotemporal depending on the model family. Following prior wildfire-management studies [50, 85, 15, 17, 6], we define the initial attack below.

Definition 2.1 (Initial Attack)**.** For a reported wildfire event i, the *initial attack* phase is the first suppression phase beginning at t_i^d, during which first-arriving firefighting resources attempt to halt or contain fire spread. A wildfire *escapes initial attack* if this first response fails and the incident requires suppression beyond the initial response capacity.

Because direct operational IA outcomes are not consistently reported in public national records, we follow prior size-based practice and define fires no larger than 10 ha as successful IA ($y_i = 0$), fires at least 50 ha as IA failure ($y_i = 1$), and exclude intermediate-size fires to reduce label ambiguity [85, 50, 17]. IA Failure prediction is then defined as:

Problem 2.2 (IA Failure Prediction)**.** Given the discovery-time input $\boldsymbol{x}_i(\mathcal{S})$ and binary IA label y_i, the task is to learn a predictor $p_i(\mathcal{S}) = f_{\boldsymbol{\theta}}\big(\boldsymbol{x}_i(\mathcal{S})\big) \approx \Pr(y_i = 1 \mid \boldsymbol{x}_i(\mathcal{S}))$, where $p_i(\mathcal{S}) \in [0, 1]$ is the predicted probability that wildfire event i escapes initial attack. The full benchmark dataset is $\mathcal{D} = \{(\boldsymbol{x}_i(\mathcal{C}), y_i)\}_{i=1}^n$. In WILDFIREIA, $\mathcal{C}$ includes FPA-FOD metadata, FIRMS/VIIRS fire-signal features, gridMET weather and fire-danger features, LANDFIRE vegetation, fuel, and topography features, OpenStreetMap access features, and WorldPop population-context features. Different choices of $\mathcal{S}$ instantiate the full-source, source-ablation, static-source, and weather-history tasks evaluated in later sections. Based on this task design, WILDFIREIA studies five research questions.

RQ1: Full-source predictability. How accurately can public discovery-time sources predict IA failure, and which input representation is most effective?

RQ2: Source necessity. Under the full-input setting, which source groups contribute non-redundant predictive signal, and which become largely redundant once other sources are available?

RQ3: Static-source value. When weather history and discovery-day satellite detections are unavailable, which static sources improve prediction beyond FPA-FOD metadata?

RQ4: Weather-history sufficiency. How much near-discovery weather and fire-danger history is needed for IA failure prediction, and does a multi-day history improve over discovery-day conditions?

RQ5: Containment-duration signal. Can the same discovery-time inputs predict containment duration, or is this post-discovery outcome driven by factors outside the benchmark input contract?

2.2 Data Sources and Collection

Table 1 summarizes the public data products used in WILDFIREIA, covering the event backbone, discovery-day fire signal, weather and fire-danger history, static landscape context, access proxies, and population context. Figure 1 illustrates how these event-centered layers are aligned around each reported wildfire before being converted into benchmark features.

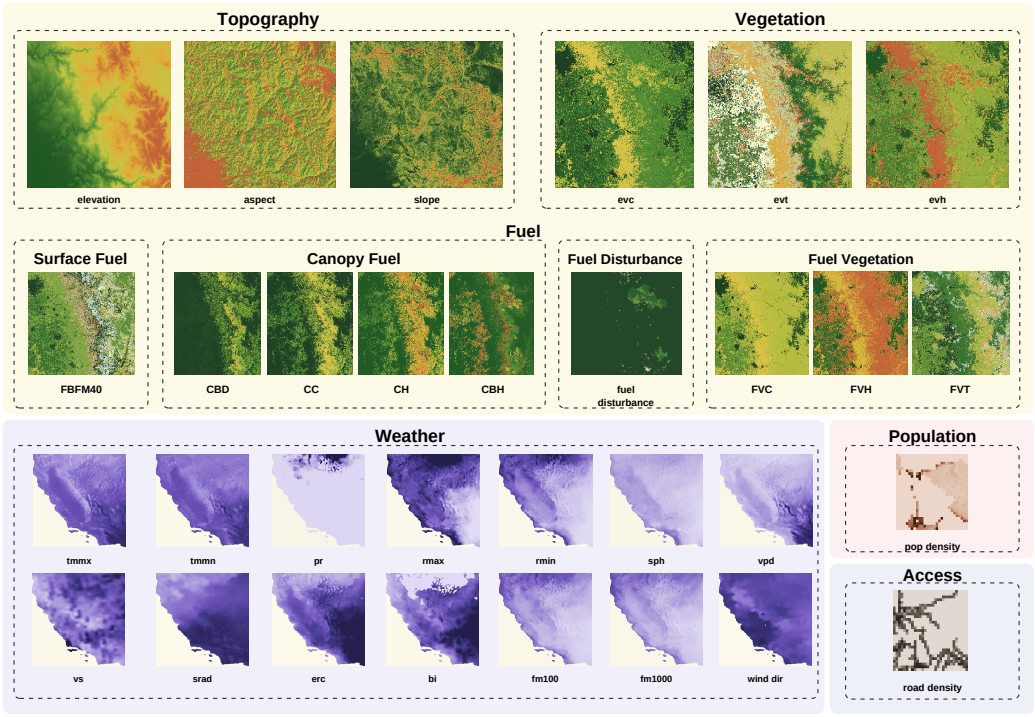

Figure 1: Examples of event-centered source layers used by WILDFIREIA. The visualization shows aligned topography, vegetation, fuel, weather, population, and access channels around a reported wildfire location. These layers are discovery-time or static inputs, not prediction targets.

Event backbone. WILDFIREIA uses FPA-FOD as the event backbone [74], retaining 38,128 naturally caused wildfires from 2016–2020 in the contiguous United States with valid identifiers, discovery dates, coordinates, and positive final sizes. This yields a rare-event task, with IA failure rates of 7.55%, 6.07%, and 8.32% in train, validation, and test.

Discovery-day fire signal. FIRMS/VIIRS provides 375 m active-fire observations used only as discovery-day thermal evidence, not as the event definition [58, 69]. We match 1,738 detections to nearby FPA-FOD events and summarize count, fire radiative power, brightness, and patch-level channels; unmatched events receive zero-valued thermal features.

Weather. gridMET provides daily meteorological and fire-danger variables through discovery day at approximately 4 km resolution [1]. For each event we extract a D-4:D discovery-window sequence covering temperature, humidity, wind, energy release component, burning index, and fuel moisture.

Topography. LANDFIRE topography provides CONUS-scale terrain layers at 30 m resolution [53]. We use elevation, slope, and aspect as fixed terrain attributes around each reported fire location.

Vegetation. LANDFIRE vegetation provides 30 m CONUS-wide landscape structure and plant-community information [53]. We use Existing Vegetation Type (EVT), Existing Vegetation Cover (EVC), and Existing Vegetation Height (EVH) to describe local vegetation composition and structure.

Fuel. LANDFIRE fuel products provide 30 m combustible-material layers over the contiguous United States [53]. We use surface fuel (FBFM40), canopy fuel variables, fuel disturbance, and fuel vegetation layers to represent persistent fuel conditions relevant to fire growth potential.

Access. OpenStreetMap provides access proxies, including drivable-road density and fire-station proximity, at both event and patch levels [41]. The benchmark uses a 2020 extract with 26.0M drivable-road geometries, 12.8M km of projected road linework, and 27,989 fire-station geometries.

Population. WorldPop provides prior-year annual population density for each fire year, using the preceding year's raster to avoid future information [78]. The annual rasters use 0.000833° cells, approximately 100 m resolution. The 2019 raster contains 430,711 $\times$ 62,976 cells.

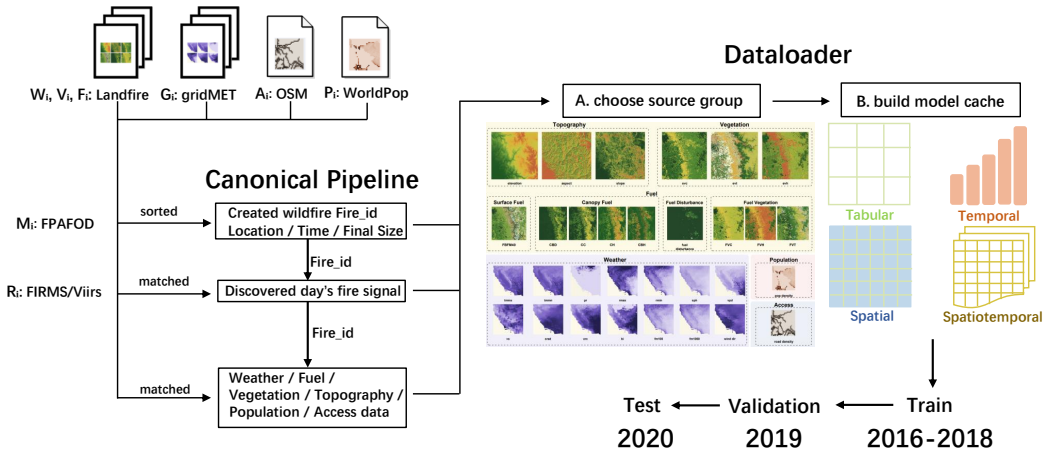

Figure 2: The full construction pipeline aligns raw data into canonicalized data, loads it into a training-ready cache, and defines the train/validation/test setting.

2.3 Canonicalization Pipeline

Figure 2 summarizes the construction pipeline used to enforce a common event unit across all sources. Each retained FPA-FOD record defines one benchmark sample with a unique `fire_id`; its discovery date, location, and fire year determine how external public sources are attached. Raw sources are aligned to these event records and canonicalized into event-level and patch-level artifacts keyed by `fire_id`. The dataloader then applies the requested source setting and discovery-time rules before converting the canonical artifacts into tabular, temporal, spatial, and spatiotemporal model caches.

Label construction. Each retained FPA-FOD record defines one benchmark sample with a unique fire_id. The event's discovery date, location, and fire year serve as the join key for attaching all external sources. Final fire size and containment timestamps are used solely for label construction and are excluded from model inputs via a forbidden-feature manifest.

Fire signal matching. FIRMS/VIIRS detections are attached to FPA-FOD events only on the reported discovery day. Each VIIRS detection is assigned to at most one nearby FPA-FOD event within the matching radius. Unmatched VIIRS detections are discarded, while FPA-FOD events without matched VIIRS detections are retained with zero-valued thermal features and a detection indicator. D+1 and later detections are excluded to avoid post-discovery fire-growth leakage.

Environmental and contextual features. We extract gridMET weather and fire-danger features through discovery day, and use LANDFIRE fuel, vegetation, and topography as static landscape context. OpenStreetMap provides access and response-proximity features, while prior-year WorldPop rasters provide population context. All features are stored as `fire_id`-indexed canonical tables.

Spatial artifacts and outputs. For spatial models, we build an event-centered 29×29 grid at 375 m resolution in EPSG:5070 around each discovery location. Raster, vector, and point sources are sampled on this grid to produce spatial and spatiotemporal caches. The canonical outputs include event tables, source-specific features, daily weather records, spatial artifacts, and a master table with fixed splits, input windows, and forbidden-feature rules.

3 Tasks, Metrics & Evaluation Protocol

In this section we introduce how WILDFIREIA evaluates IA failure prediction. All tasks share the same event unit, label rule, discovery-time input constraint, and leakage-control policy, while varying the source subset, weather window, representation family, or prediction target. We then specify metrics for rare-event classification and auxiliary containment-duration prediction.

3.1 Evaluation Tasks

All tasks use the label y_i, source set $\mathcal{C}$, and input notation $\boldsymbol{x}_i(\mathcal{S})$ from Problem 2.2, and all inputs follow the discovery-time leakage-control rule in Section 2.1.

Task 1: Full-source predictability. This task evaluates whether the complete public discovery-time source stack can predict initial attack failure. Each model receives $\boldsymbol{x}_i(\mathcal{C})$ and predicts

$$p_i^{\mathrm{IA}} = f_{\boldsymbol{\theta}}\big(\boldsymbol{x}_i(\mathcal{C})\big),$$

where $p_i^{\mathrm{IA}} \in [0,1]$ is the predicted probability that event i escapes initial attack.

Task 2: Source necessity under full input. This task measures whether each source group contributes non-redundant information when all other sources are available. For each source group $\mathcal{Q} \in \mathcal{C} \setminus \{\mathcal{M}\}$, the model receives $\boldsymbol{x}_i(\mathcal{C} \setminus \{\mathcal{Q}\})$ and predicts y_i. A large performance drop after removing $\mathcal{Q}$ indicates that the removed source contains information not recovered by the remaining source groups.

Task 3: Static-source value without dynamic observations. This task evaluates which static source remains useful when dynamic observations are unavailable. We remove FIRMS/VIIRS and weather by excluding $\mathcal{R}$ and $\mathcal{W}$, and compare metadata plus one static source group, $\boldsymbol{x}_i(\{\mathcal{M}, \mathcal{Q}\})$, for each $\mathcal{Q} \in \{\mathcal{V}, \mathcal{F}, \mathcal{T}, \mathcal{A}, \mathcal{P}\}$. We also evaluate $\boldsymbol{x}_i(\{\mathcal{M}\})$ as the metadata-only lower bound.

Task 4: Weather-history sufficiency. This task evaluates how much retrospective weather history is needed for initial attack failure prediction. Let $\ell \in \{1,2,3,4,5\}$ denote the number of days in the weather window ending on discovery day t_i^d, and let $\mathcal{W}^{(\ell)} \subseteq \mathcal{W}$ denote the ℓ-day gridMET weather and fire-danger history. Each model receives $\boldsymbol{x}_i(\{\mathcal{M}, \mathcal{W}^{(\ell)}\})$ and predicts y_i. Comparing performance across values of ℓ tests whether longer weather histories improve prediction beyond weather near discovery time.

Task 5: Auxiliary containment-duration prediction. This task tests whether the same discovery-time public inputs can predict containment duration. For events with valid timestamps, let t_i^c denote containment time; we define containment duration $h_i = t_i^c - t_i^d$ and log-transformed target $y_i^{\mathrm{TTC}} = \log(1 + h_i)$. The model receives $\boldsymbol{x}_i(\mathcal{C})$ and predicts

$$\hat{y}_i^{\mathrm{TTC}} = g_{\boldsymbol{\phi}}\big(\boldsymbol{x}_i(\mathcal{C})\big), \qquad \hat{h}_i = \exp(\hat{y}_i^{\mathrm{TTC}}) - 1,$$

where $g_{\boldsymbol{\phi}}$ is a predictor parameterized by vector $\boldsymbol{\phi}$. This auxiliary task is not used to define initial attack failure. It serves as a stress test of how much post-discovery suppression outcome can be explained by discovery-time public data.

3.2 Evaluation Metrics

IA failure metrics. For IA failure prediction, we use **AUPRC** as the primary metric because escaped fires are rare and the benchmark emphasizes ranking high-risk events near the top. We also report **AUROC**, **Recall@5%**, **F1**, **Brier score**, and **ECE**. These metrics measure ranking quality, early-warning recall, thresholded classification performance, probability error, and calibration.

Containment-duration metrics. For auxiliary containment-duration prediction, we use **MAE** in hours as the primary metric after converting log-space predictions back to containment hours. We also report **RMSE**, **MedianAE**, **log-space MAE**, R^2, and **Spearman correlation**. These metrics measure large-error sensitivity, typical-event error, training-scale error, explained variance, and ranking consistency across containment durations.

4 Experimental Setup

All experiments use the same canonical WILDFIREIA artifacts. Each sample is one FPA-FOD natural wildfire event indexed by `fire_id`. We use a chronological split, with 2016–2018 fires for training, 2019 for validation, and 2020 for testing. Across all experiments, we keep the event set, labels, forbidden-feature list, and five-seed protocol fixed, so that differences reflect source choices, representations, or model families rather than changes in benchmark construction.

4.1 Model-Ready Representations

WILDFIREIA compares model families under the same event-level IA failure prediction contract. After public sources are aligned by `fire_id`, the dataloader applies the requested task and source setting, removes forbidden fields, fits preprocessing statistics on the training split only, and exports model-ready caches. This ensures that models use the same events, labels, splits, and leakage policy,

Table 2: Model-ready representations and baseline families in WILDFIREIA.

Family	Input shape	Models
Tabular	$N \times F$	logistic regression, XGBoost, MLP
Temporal	$N \times T \times F_t$ plus $N \times F_s$	GRU, TCN, Transformer
Spatial	$N \times C \times 29 \times 29$	ResNet18-UNet, ResNet50-UNet, Swin-UNet, SegFormer
Spatiotemporal	$N \times T \times C \times 29 \times 29$	ConvLSTM, ConvGRU, ResNet3D, PredRNN-V2, UTAE, SwinLSTM

Table 3: Full-Source Predictability leaderboard. Values are percentages and shown as mean±std. AUPRC is the primary metric; AUROC and Recall@5% evaluate ranking, F1 evaluates a validation-selected operating point, and Brier/ECE evaluate calibration. Best values are bolded.

Family	Model	AUPRC↑	AUROC↑	Recall@5%↑	F1↑	Brier↓	ECE↓
Tabular	LogisticRegression	$43.9_{\pm 0.0}$	$81.5_{\pm 0.0}$	$32.9_{\pm 0.0}$	$42.5_{\pm 0.0}$	$13.6_{\pm 0.0}$	$15.8_{\pm 0.0}$
	XGBoost	$\mathbf{53.3}_{\pm 0.3}$	$\mathbf{87.1}_{\pm 0.1}$	$\mathbf{38.0}_{\pm 0.6}$	$\mathbf{49.5}_{\pm 0.6}$	$9.1_{\pm 0.1}$	$13.5_{\pm 0.1}$
	MLP	$50.5_{\pm 0.5}$	$83.4_{\pm 0.5}$	$36.7_{\pm 0.3}$	$47.1_{\pm 0.6}$	$6.1_{\pm 0.1}$	$5.9_{\pm 0.7}$
Temporal	GRU	$49.5_{\pm 0.1}$	$82.9_{\pm 0.4}$	$36.3_{\pm 0.6}$	$46.0_{\pm 0.8}$	$5.7_{\pm 0.0}$	$\mathbf{0.9}_{\pm 0.3}$
	TCN	$50.4_{\pm 0.3}$	$83.7_{\pm 0.5}$	$36.4_{\pm 0.3}$	$46.7_{\pm 0.6}$	$5.8_{\pm 0.0}$	$3.5_{\pm 0.6}$
	Transformer	$50.6_{\pm 0.3}$	$84.0_{\pm 0.4}$	$36.5_{\pm 0.3}$	$46.8_{\pm 0.6}$	$5.6_{\pm 0.0}$	$0.9_{\pm 0.4}$
Spatial	ResNet18-UNet	$50.2_{\pm 1.7}$	$84.8_{\pm 0.7}$	$35.4_{\pm 1.1}$	$47.2_{\pm 0.8}$	$5.8_{\pm 0.3}$	$4.3_{\pm 1.4}$
	ResNet50-UNet	$49.4_{\pm 0.6}$	$84.0_{\pm 0.9}$	$34.9_{\pm 1.2}$	$44.8_{\pm 1.5}$	$5.8_{\pm 0.1}$	$3.3_{\pm 0.3}$
	Swin-UNet	$51.1_{\pm 1.2}$	$85.3_{\pm 0.6}$	$36.5_{\pm 1.8}$	$46.7_{\pm 1.9}$	$5.6_{\pm 0.1}$	$2.6_{\pm 0.6}$
	SegFormer	$50.6_{\pm 0.5}$	$84.9_{\pm 0.9}$	$35.8_{\pm 0.4}$	$46.6_{\pm 1.2}$	$5.7_{\pm 0.1}$	$2.3_{\pm 0.3}$
Spatiotemporal	ConvLSTM	$38.9_{\pm 1.7}$	$83.8_{\pm 0.3}$	$29.7_{\pm 2.0}$	$40.4_{\pm 1.0}$	$6.3_{\pm 0.1}$	$3.3_{\pm 0.4}$
	ConvGRU	$41.2_{\pm 1.6}$	$84.1_{\pm 0.6}$	$31.6_{\pm 0.9}$	$41.6_{\pm 1.1}$	$6.2_{\pm 0.1}$	$3.5_{\pm 0.2}$
	ResNet3D	$51.3_{\pm 1.7}$	$84.5_{\pm 2.0}$	$36.5_{\pm 1.5}$	$46.7_{\pm 1.1}$	$5.7_{\pm 0.1}$	$3.5_{\pm 1.3}$
	PredRNN-V2	$39.6_{\pm 1.6}$	$83.4_{\pm 0.4}$	$30.8_{\pm 1.1}$	$41.8_{\pm 2.0}$	$6.3_{\pm 0.1}$	$2.7_{\pm 0.5}$
	UTAE	$51.3_{\pm 1.5}$	$85.1_{\pm 1.2}$	$36.2_{\pm 0.8}$	$46.6_{\pm 1.1}$	$5.7_{\pm 0.2}$	$3.3_{\pm 1.2}$
	SwinLSTM	$51.7_{\pm 1.2}$	$85.7_{\pm 0.9}$	$37.3_{\pm 1.3}$	$47.1_{\pm 2.3}$	$\mathbf{5.6}_{\pm 0.1}$	$2.4_{\pm 0.7}$

while differing only in input representation. As summarized in Table 2, WILDFIREIA provides four views: a tabular view that aggregates each event into structured features, a temporal view that preserves the $D-4{:}D$ weather and fire-danger history, a spatial view that represents each fire as an event-centered 29×29 patch, and a spatiotemporal view that combines patch-level context with short weather dynamics. All views produce the same scalar event-level output, either IA failure probability or auxiliary containment duration, rather than a fire mask, perimeter, or spread trajectory.

4.2 Baseline Model Suite

We evaluate 16 baselines across tabular, temporal, spatial, and spatiotemporal families. The suite covers structured-data models, recurrent and attention-based sequence encoders, convolutional and transformer-based spatial encoders, and recurrent or 3D spatiotemporal architectures [21, 23, 8, 80, 42, 66, 56, 16, 84, 71, 81, 34, 77]. Segmentation-style architectures are adapted by replacing dense prediction heads with global pooling and a scalar output, so every model predicts either initial attack failure probability or auxiliary containment duration.

4.3 Training and Leakage Control

All preprocessing is fit on the training split only and then fixed for validation and test events. Before training, the dataloader removes forbidden fields, including final fire size, containment timestamps, MTBS identifiers, post-discovery detections, and label-derived variables. Discovery-day FIRMS/VIIRS features are allowed, but later detections are excluded; unmatched events are retained with zero-valued thermal features and a detection indicator.

5 Experiments

We organize the experiments around the five research questions defined in Section 2.1. All experiments share the same event unit, chronological split, leakage policy, and evaluation protocol, making results comparable across model families, source combinations, weather windows, and prediction targets.

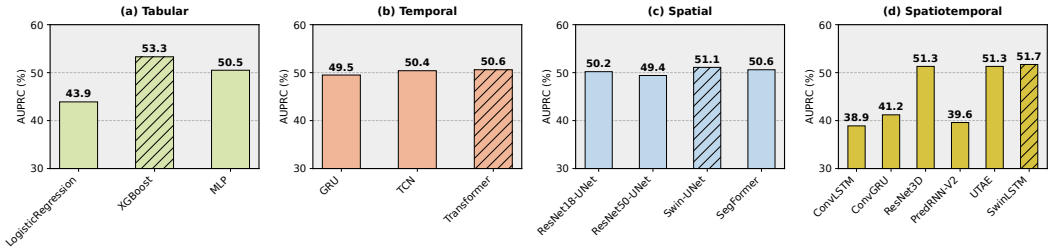

Figure 3: Full-Source Initial Attack Predictability leaderboard.

Table 4: Source Necessity under Full Input for initial attack escape prediction. Each cell is AUPRC in percent after removing one source from the full input, reported as mean±standard deviation over five random seeds. The final column reports the unablated full-input result.

Model	w/o FIRMS	w/o Weather	w/o Vegetation	w/o Fuel	w/o Topography	w/o Access	w/o Population	Full
XGBoost	$38.3_{\pm 0.5}$	$50.5_{\pm 0.5}$	$\mathbf{53.6}_{\pm 0.3}$	$52.1_{\pm 0.3}$	$53.1_{\pm 0.2}$	$53.6_{\pm 0.4}$	$51.2_{\pm 0.4}$	$53.3_{\pm 0.3}$
Transformer	$32.3_{\pm 0.2}$	$49.2_{\pm 0.3}$	$\mathbf{51.4}_{\pm 0.3}$	$49.7_{\pm 0.3}$	$50.7_{\pm 0.2}$	$50.2_{\pm 0.5}$	$50.3_{\pm 0.6}$	$50.6_{\pm 0.3}$
Swin-UNet	$30.4_{\pm 1.0}$	$48.8_{\pm 1.2}$	$48.2_{\pm 1.5}$	$49.5_{\pm 1.1}$	$50.6_{\pm 0.9}$	$46.2_{\pm 1.1}$	$50.0_{\pm 1.4}$	$\mathbf{51.1}_{\pm 1.2}$
SwinLSTM	$33.4_{\pm 1.1}$	$48.1_{\pm 0.8}$	$51.3_{\pm 0.7}$	$\mathbf{52.2}_{\pm 0.4}$	$51.0_{\pm 1.0}$	$50.6_{\pm 0.9}$	$51.6_{\pm 1.1}$	$51.7_{\pm 1.2}$

5.1 Full-Source Predictability

To answer **RQ1**, we evaluate the full-source setting using all discovery-time environmental and contextual inputs across tabular, temporal, spatial, and spatiotemporal model families. Table 3 reports the full metric suite over five random seeds, and Figure 3 visualizes the AUPRC leaderboard by representation family. We find that: (1) The full-source setting contains a clear predictive signal: all major representation families substantially outperform a trivial rare-event baseline, showing that initial attack failure can be predicted from public discovery-time and static contextual variables. (2) The task behaves primarily as event-level risk prediction rather than dense spatial forecasting. Higher-capacity spatial and spatiotemporal models do not automatically outperform structured event-level models, and the best results are obtained by XGBoost. (3) Patch-based models are still competitive, indicating that local landscape and discovery-day spatial context are useful, but they do not replace aggregated event-level summaries. At the same time, the best Recall@5% remains limited, showing that public discovery-time inputs do not fully determine suppression outcome.

5.2 Source Necessity under Full Input.

To answer **RQ2**, we perform leave-one-source-out ablations and compare each setting with the full-source protocol. This measures marginal necessity: a large AUPRC drop indicates that the removed source provides signal not recovered by the remaining inputs. Table 4 shows that FIRMS/VIIRS is the least redundant source, as removing discovery-day thermal evidence causes the largest degradation across representative models. Weather provides a smaller but consistent gain, suggesting partial overlap with other contextual variables. Vegetation, fuel, topography, access, and population have marginal or even slightly negative effects in the full-input setting, indicating that their information is partly covered once FIRMS/VIIRS, weather, metadata, and other contextual sources are available.

5.3 Static-source value without dynamic observations.

To answer **RQ3**, we remove both dynamic sources and evaluate two conditions: FPA-FOD metadata alone as the lower bound, and FPA-FOD metadata combined with one static source group at a time. This fallback setting measures which static information remains useful without weather history or discovery-day satellite evidence. Table 5 shows that fuel is the strongest static source across representative models, suggesting that combustible landscape structure provides meaningful back-

Table 5: Static-source ablation. AUPRC is reported in percentage over five seeds.

Model	Vegetation	Fuel	Topography	Access	Population	FPA-FOD
XGBoost	$25.6_{\pm 0.2}$	$\mathbf{30.3}_{\pm 0.6}$	$25.6_{\pm 0.5}$	$24.2_{\pm 0.3}$	$26.7_{\pm 0.4}$	$23.6_{\pm 0.5}$
Transformer	$25.3_{\pm 0.6}$	$\mathbf{28.8}_{\pm 0.3}$	$24.6_{\pm 0.2}$	$23.9_{\pm 0.5}$	$20.9_{\pm 0.5}$	$22.8_{\pm 0.4}$
Swin-UNet	$22.1_{\pm 2.4}$	$\mathbf{23.3}_{\pm 1.7}$	$21.2_{\pm 3.0}$	$20.0_{\pm 1.6}$	$17.0_{\pm 0.8}$	$15.9_{\pm 0.7}$
SwinLSTM	$26.4_{\pm 1.2}$	$\mathbf{26.7}_{\pm 2.1}$	$21.1_{\pm 0.9}$	$20.8_{\pm 1.2}$	$15.5_{\pm 1.0}$	$14.5_{\pm 1.3}$

Table 6: Containment-Duration prediction over five random seeds. Models predict log containment hours; MAE, RMSE, and MedianAE are reported after converting predictions back to hours. Lower is better for error metrics, and higher is better for R^2 and Spearman correlation.

Family	Model	MAE h↓	RMSE h↓	MedianAE h↓	log MAE↓	R^2↑	Spearman↑
Tabular	RidgeRegression	$39.3_{\pm 0.0}$	$129.1_{\pm 0.0}$	$8.4_{\pm 0.0}$	$1.319_{\pm 0.000}$	$0.058_{\pm 0.000}$	$0.345_{\pm 0.000}$
	XGBoost	$\mathbf{35.1}_{\pm 0.0}$	$\mathbf{121.3}_{\pm 0.2}$	$8.2_{\pm 0.1}$	$\mathbf{1.197}_{\pm 0.002}$	$0.172_{\pm 0.004}$	$\mathbf{0.411}_{\pm 0.002}$
	MLP	$35.5_{\pm 0.3}$	$125.6_{\pm 0.8}$	$\mathbf{4.9}_{\pm 0.2}$	$1.209_{\pm 0.011}$	$0.176_{\pm 0.016}$	$0.396_{\pm 0.005}$
Temporal	GRU	$36.4_{\pm 0.3}$	$126.2_{\pm 3.1}$	$7.3_{\pm 0.1}$	$1.222_{\pm 0.006}$	$\mathbf{0.177}_{\pm 0.011}$	$0.375_{\pm 0.007}$
	TCN	$36.1_{\pm 0.3}$	$124.7_{\pm 0.7}$	$6.6_{\pm 0.3}$	$1.226_{\pm 0.003}$	$0.157_{\pm 0.007}$	$0.376_{\pm 0.005}$
	Transformer	$36.0_{\pm 0.3}$	$124.1_{\pm 1.3}$	$6.9_{\pm 0.5}$	$1.216_{\pm 0.010}$	$0.168_{\pm 0.015}$	$0.377_{\pm 0.007}$
Spatial	ResNet18-UNet	$36.0_{\pm 0.3}$	$125.2_{\pm 0.4}$	$8.1_{\pm 0.1}$	$1.231_{\pm 0.006}$	$0.150_{\pm 0.013}$	$0.365_{\pm 0.011}$
	ResNet50-UNet	$37.2_{\pm 2.3}$	$128.5_{\pm 2.6}$	$8.4_{\pm 0.5}$	$1.263_{\pm 0.014}$	$0.110_{\pm 0.018}$	$0.338_{\pm 0.009}$
	Swin-UNet	$36.9_{\pm 0.4}$	$128.7_{\pm 6.0}$	$8.3_{\pm 0.5}$	$1.274_{\pm 0.026}$	$0.114_{\pm 0.028}$	$0.333_{\pm 0.024}$
	SegFormer	$36.9_{\pm 0.6}$	$126.9_{\pm 2.8}$	$8.3_{\pm 0.4}$	$1.263_{\pm 0.021}$	$0.106_{\pm 0.025}$	$0.337_{\pm 0.019}$
Spatiotemporal	ConvLSTM	$36.5_{\pm 0.1}$	$125.8_{\pm 0.5}$	$7.8_{\pm 0.4}$	$1.251_{\pm 0.009}$	$0.136_{\pm 0.013}$	$0.341_{\pm 0.008}$
	ConvGRU	$36.4_{\pm 0.1}$	$125.5_{\pm 0.4}$	$7.8_{\pm 0.2}$	$1.243_{\pm 0.004}$	$0.150_{\pm 0.008}$	$0.353_{\pm 0.005}$
	ResNet3D	$37.5_{\pm 2.3}$	$127.4_{\pm 1.2}$	$8.4_{\pm 0.6}$	$1.266_{\pm 0.042}$	$0.126_{\pm 0.055}$	$0.356_{\pm 0.022}$
	PredRNN-V2	$36.4_{\pm 0.1}$	$125.6_{\pm 0.4}$	$7.9_{\pm 0.3}$	$1.243_{\pm 0.008}$	$0.145_{\pm 0.009}$	$0.353_{\pm 0.008}$
	UTAE	$37.6_{\pm 2.1}$	$131.7_{\pm 3.8}$	$6.6_{\pm 1.9}$	$1.300_{\pm 0.060}$	$0.080_{\pm 0.114}$	$0.309_{\pm 0.019}$
	SwinLSTM	$36.3_{\pm 0.2}$	$125.6_{\pm 0.7}$	$7.8_{\pm 0.6}$	$1.231_{\pm 0.014}$	$0.151_{\pm 0.016}$	$0.365_{\pm 0.010}$

ground risk signal. The metadata-only condition confirms the lower bound: all static sources provide a lift over using metadata alone, with fuel showing the largest and most consistent gain. However, static sources remain insufficient on their own: vegetation, topography, access, and population are weaker and more model-dependent, indicating that static context can support early risk assessment but cannot replace current fire signal or near-discovery weather observations.

5.4 Weather-History Sufficiency.

To answer **RQ4**, we vary the weather and fire-danger window from discovery day to the full $D-4{:}D$ history while keeping the initial attack failure target fixed. Table 7 shows that longer weather histories do not consistently improve prediction: the best or near-best AUPRC values usually occur with one or two days of weather, while adding three to five days provides little additional benefit. This suggests that, under the current event-level target and gridMET feature design, the most useful weather signal is concentrated near discovery time, even though longer antecedent conditions may remain physically relevant.

Table 7: Weather-history sufficiency for initial attack escape prediction. Each cell is AUPRC in percentage. Best value in each row is bolded.

Model	1d	2d	3d	4d	5d
XGBoost	$\mathbf{20.1}_{\pm 0.2}$	$18.8_{\pm 0.3}$	$19.0_{\pm 0.2}$	$19.1_{\pm 0.2}$	$19.0_{\pm 0.6}$
Transformer	$25.1_{\pm 0.2}$	$\mathbf{25.3}_{\pm 0.3}$	$24.7_{\pm 0.6}$	$24.2_{\pm 0.8}$	$23.8_{\pm 0.8}$
Swin-UNet	$22.4_{\pm 0.6}$	$\mathbf{22.8}_{\pm 1.0}$	$22.7_{\pm 1.2}$	$21.8_{\pm 1.1}$	$21.6_{\pm 2.7}$
SwinLSTM	$23.1_{\pm 1.3}$	$\mathbf{23.7}_{\pm 0.6}$	$22.7_{\pm 1.7}$	$23.1_{\pm 1.4}$	$23.0_{\pm 1.4}$

5.5 Auxiliary Containment-Duration Signal.

To answer **RQ5**, we test the boundary of the discovery-time benchmark contract by replacing the initial attack label with log containment hours. Table 6 shows that XGBoost remains the strongest baseline, indicating that public discovery-time inputs contain some severity signal. However, the low R^2 suggests that containment duration is only weakly explained by these inputs and is largely shaped by post-discovery resources, tactics, weather evolution, and reporting practices. This reinforces our main design choice: WILDFIREIA is best suited for early event-level triage, while containment duration serves as an auxiliary stress test of the benchmark's limits.

6 Conclusion

We introduce WILDFIREIA, a reproducible AI benchmark for event-level wildfire initial attack failure prediction from public U.S. national environmental data. WILDFIREIA fixes the event unit, discovery-time input contract, leakage policy, temporal split, metrics, and model-ready representations, enabling fair comparison across tabular, temporal, spatial, and spatiotemporal models. Experiments show that discovery-time public data contains useful early-risk signals, while also revealing source-specific contributions and the limits of containment-duration prediction.

References

[1] John T. Abatzoglou. Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1):121–131, 2013.

[2] John T. Abatzoglou and A. Park Williams. Impact of anthropogenic climate change on wildfire across western us forests. *Proceedings of the National Academy of Sciences*, 113(42):11770–11775, 2016.

[3] Anonymous. Firepred: A hybrid multi-temporal convolutional neural network model for wildfire spread prediction. *Ecological Informatics*, 2023.

[4] Anonymous. Application of explainable artificial intelligence in predicting wildfire spread: An aspp-enabled cnn approach. *IEEE Geoscience and Remote Sensing Letters*, 2024. DOI: 10.1109/LGRS.2024.3417624.

[5] Anonymous. Wildfire spread prediction in north america using satellite imagery and vision transformer. In *Proceedings of the 2024 IEEE Conference on Artificial Intelligence*, 2024. DOI: 10.1109/CAI59869.2024.00278.

[6] M. Cecilia Arienti, Steven G. Cumming, and Stan Boutin. Empirical models of forest fire initial attack success probabilities: the effects of fuels, anthropogenic linear features, fire weather, and management. *Canadian Journal of Forest Research*, 36(12):3155–3166, 2006.

[7] Cong Bai, Feng Sun, Jinglin Zhang, Yi Song, and Shengyong Chen. Rainformer: Features extraction balanced network for radar-based precipitation nowcasting. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[8] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[9] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[10] Jennifer K. Balch, Bethany A. Bradley, John T. Abatzoglou, R. Chelsea Nagy, Emily J. Fusco, and Adam L. Mahood. Human-started wildfires expand the fire niche across the united states. *Proceedings of the National Academy of Sciences*, 114(11):2946–2951, 2017.

[11] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. In *International Conference on Learning Representations (ICLR)*, 2016. Introduces convolutional GRU variants.

[12] Shashank Bhardwaj. Predicting the containment time of california wildfires using machine learning, 2025. arXiv preprint.

[13] Leo Breiman. Random forests. *Machine Learning*, 2001.

[14] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 1950.

[15] Kristy Butler, Elena Tartaglia, Jason Rennie, Stephen Deutsch, and Nick McCarthy. An application-specific approach to modelling the probability of unsuccessful initial attack on wildfires in victoria, australia, 2025. bioRxiv preprint.

[16] Hu Cao, Yueyue Wang, Joy Chen, Dong Jiang, Xiaoqing Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[17] Adrián Cardil, Adrián Jiménez-Ruano, Santiago Monedero, Phillip SeLegue, Macarena Ortega, Raúl Quilez, Jeff Fuentes, Geoff Marshall, Robert Clark, Tim Chavez, et al. Assessing the suppression difficulty of wildland fires for initial attack response. *International Journal of Wildland Fire*, 34:WF24160, 2025. doi:10.1071/WF24160.

[18] Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Yan Ye, Xiang Xinguang, and Wen Gao. Mau: A motion-aware unit for video prediction and beyond. In *Advances in Neural Information Processing Systems*, 2021.

[19] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[20] Mengxuan Chen, Guowen Li, Fang Wang, Runmin Dong, Juepeng Zheng, Ziheng Zou, Jinxiao Zhang, and Haohuan Fu. Seasonbench-ea: A multi-source benchmark for seasonal prediction and numerical model post-processing in east asia, 2025. Benchmark paper.

[21] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[22] Santos Daniel Chicas and Jonas Østergaard Nielsen. Who are the actors and what are the factors that are used in models to map forest fire susceptibility? a systematic review. *Natural Hazards*, 114:2417–2434, 2022.

[23] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014.

[24] Janice L. Coen et al. Wrf-fire: Coupled weather-wildland fire modeling with the weather research and forecasting model. *Geoscientific Model Development*, 2011.

[25] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 1995.

[26] David R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1958.

[27] Weizhi Deng, Jun Wang, Meng Zhou, et al. Fires reverse progress toward ozone air quality standards in the united states. *Science*, 392(6802):1088–1092, 2026.

[28] Philip E. Dennison, Simon C. Brewer, James D. Arnold, and Max A. Moritz. Large wildfire trends in the western united states, 1984–2011. *Geophysical Research Letters*, 41(8):2928–2933, 2014.

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[30] Jean-Baptiste Filippi et al. Forefire: Open source framework for wildland fire spread modelling. In *Advances in Forest Fire Research*. 2014.

[31] Mark A. Finney. Farsite: Fire area simulator–model development and evaluation. Technical Report RMRS-RP-4, USDA Forest Service, Rocky Mountain Research Station, 1998.

[32] Mark A. Finney. An overview of flammap fire modeling capabilities. In *Fuels Management—How to Measure Success: Conference Proceedings*, 2006.

[33] Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. In *Advances in Neural Information Processing Systems*, 2022.

[34] Vivien Sainte Fare Garnot, Loïc Landrieu, Sébastien Giordano, and Nesrine Chehata. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[35] Ronald Gelaro, Will McCarty, Max J. Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A. Randles, Anton Darmenov, Michael G. Bosilovich, Rolf Reichle, et al. The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of Climate*, 2017.

[36] Sean Gerard, Yuwen Johnson, and Jordan Malof. Wildfirespreadts: A dataset of multi-modal time series for wildfire spread prediction. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

[37] Rafik Ghali and Moulay A. Akhloufi. Deep learning approaches for wildland fires remote sensing: Classification, detection, and segmentation. *Remote Sensing*, 15(7):1821, 2023.

[38] Rafik Ghali and Moulay A. Akhloufi. Deep learning approaches for wildland fires using satellite remote sensing data: Detection, mapping, and prediction. *Fire*, 6(5):192, 2023.

[39] Louis Giglio, Wilfrid Schroeder, and Christopher O. Justice. Collection 6 modis active fire detection algorithm and fire products. *Remote Sensing of Environment*, 2016.

[40] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[41] Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[43] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Cécile Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 2020.

[44] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[45] Fantine Huot et al. Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. DOI: 10.1109/TGRS.2022.3192974.

[46] W. Matt Jolly, Mark A. Cochrane, Patrick H. Freeborn, Zachary A. Holden, Timothy J. Brown, Grant J. Williamson, and David M. J. S. Bowman. Climate-induced variations in global wildfire danger from 1979 to 2013. *Nature Communications*, 6:7537, 2015.

[47] Hari Katuwal, Michael S. Hand, Matthew P. Thompson, Crystal Stonesifer, and David E. Calkin. Predict and attack (or don't): An econometric approach to large wildfire early detection and suppression effectiveness. In *Agricultural and Applied Economics Association Annual Meeting*, 2018.

[48] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 2017.

[49] Rey Koki, Michael McCabe, Josh Myers-Dean, Dhruv Kedar, Annabel Wade, Jebb Q. Stewart, Christina Kumler-Bonfanti, and Jed Brown. Smokeviz: A large-scale satellite dataset for wildfire smoke detection and segmentation, 2025. Dataset paper.

[50] Kennedy Korkola, Melanie Wheatley, Jennifer Beverly, Patrick M. A. James, and Mike Wotton. A comparative analysis of wildfire initial attack containment objectives and modelling strategies in ontario, canada. *International Journal of Wildland Fire*, 33:WF24104, 2024. doi:10.1071/WF24104.

[51] Taha Lahrichi, Nils Baran, Matteo Congedo, Rémi Cresson, Grégoire Mialon, Clément Mallet, Jérémie Dérappe, David Defrance, Jordan Malof, et al. Wsts+: A benchmark for spatiotemporal wildfire spread prediction and strong baselines. *arXiv preprint arXiv:2502.12003*, 2025.

[52] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

[53] LANDFIRE. LANDFIRE: Landscape fire and resource management planning tools. `https://www.landfire.gov/`, 2026. Accessed 2026.

[54] Yohan Lee, Jeremy S. Fried, Heidi J. Albers, and Robert G. Haight. Deploying initial attack resources for wildfire suppression: Spatial coordination, budget constraints, and capacity constraints. *Canadian Journal of Forest Research*, 43(1):56–65, 2013.

[55] Yanzhi Li, Lubo Wang, Keqiu Li, Die Zuo, Guohui Li, Qing Guo, Zumin Wang, Feng Zhang, Changqing Ji, Manyu Wang, and Di Lin. Sim2real-fire: A multi-modal simulation dataset for forecast and backtracking of real-world forest fire, 2024. Datasets and Benchmarks Track paper.

[56] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[57] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[58] NASA. Fire information for resource management system (firms). `https://firms.modaps.eosdis.nasa.gov/`, 2026. Accessed 2026-02-13.

[59] Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, and Pierre Gentine. Chaosbench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction, 2024. Benchmark paper.

[60] National Interagency Fire Center. National fire news. `https://www.nifc.gov/fire-information/nfn`, 2026. Accessed: 2026-06-07.

[61] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew Lee, Matthias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. In *arXiv preprint arXiv:1804.03999*, 2018.

[62] Yavar Pourmohamad, John T. Abatzoglou, Erin J. Belval, Erica Fleishman, Karen C. Short, Matthew C. Reeves, Nicholas Nauslar, Philip E. Higuera, Eric Henderson, Sawyer Ball, Amir AghaKouchak, Jeffrey P. Prestemon, Julia Olszewski, and Mojtaba Sadegh. Physical, social, and biological attributes for improved understanding and prediction of wildfires: Fpa fod-attributes dataset. *Earth System Science Data*, 16:3045–3060, 2024.

[63] Douglas Radke, Andrew Hess, and Dustin Ellsworth. Firecast: Leveraging satellites and weather data for wildfire spread prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (Workshop/Track as published)*, 2019.

[64] Eghbal Rashidi, Hugh R. Medal, and Aaron Hoskins. An attacker-defender model for analyzing the vulnerability of initial attack in wildfire suppression. *Naval Research Logistics*, 65(6):449–464, 2018.

[65] Francisco Rodríguez y Silva, Christopher D. O'Connor, Matthew P. Thompson, Juan Ramón Molina Martínez, and David E. Calkin. Modelling suppression difficulty: Current and future applications. *International Journal of Wildland Fire*, 29(8):739–751, 2020.

[66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[67] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 1986.

[68] Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, Kumar Ankur, Aman Gupta, et al. Prithvi wxc: Foundation model for weather and climate. *arXiv preprint arXiv:2409.13598*, 2024.

[69] Wilfrid Schroeder, Paulo Oliva, Louis Giglio, and Ivan A. Csiszar. The new viirs 375 m active fire detection data product: Algorithm description and initial assessment. *Remote Sensing of Environment*, 2014.

[70] Yujie Shen et al. Developing risk assessment framework for wildfire in the united states. *Journal of Sustainable Aviation and Smart Systems*, 2023.

[71] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, 2015.

[72] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. In *Advances in Neural Information Processing Systems*, 2017.

[73] Assaf Shmuel and Eyal Heifetz. Global wildfire susceptibility mapping based on machine learning models. *Forests*, 13(7):1050, 2022.

[74] Karen C. Short. Spatial wildfire occurrence data for the united states, 1992–2020 [fpa-fod]. USDA Forest Service Research Data Archive, 2022. FPA-FOD wildfire occurrence database.

[75] Samriddhi Singla, Ayan Mukhopadhyay, Michael Wilbur, Tina Diao, Vinayak Gajjewar, Ahmed Eldawy, Mykel Kochenderfer, Ross Shachter, and Abhishek Dubey. Wildfiredb: An open-source dataset connecting wildfire spread with relevant determinants. In *NeurIPS Datasets and Benchmarks Track*, 2021.

[76] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[77] Song Tang, Chuang Li, Pu Zhang, and RongNian Tang. Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[78] Andrew J. Tatem. Worldpop, open data for spatial demography. *Scientific Data*, 4:170004, 2017.

[79] USGS and USDA Forest Service. Monitoring trends in burn severity (mtbs). `https://www.mtbs.gov/`, 2026. Accessed 2026-02-13.

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[81] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S. Yu, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[82] Anthony L. Westerling, Hugo G. Hidalgo, Daniel R. Cayan, and Thomas W. Swetnam. Warming and earlier spring increase western u.s. forest wildfire activity. *Science*, 313(5789):940–943, 2006.

[83] Hao Wu, Yuxuan Liang, Wei Xiong, Zhengyang Zhou, Wei Huang, Shilong Wang, and Kun Wang. Earthfarseer: Versatile spatio-temporal dynamical systems modeling in one model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[84] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, 2021.

[85] Yiqing Xu, Kaiwen Zhou, and Fuquan Zhang. Modeling wildfire initial attack success rate based on machine learning in liangshan, china. *Forests*, 14(4):740, 2023. doi:10.3390/f14040740.

[86] Yu Zhao, Sebastian Gerard, and Yifang Ban. Ts-satfire: A multi-task satellite image time-series dataset for wildfire detection and prediction. *Scientific Data*, 12(1), 2025. DOI: 10.1038/s41597-025-06271-3.

[87] Sixiao Zheng, Jiachen Lu, Hanqing Zhao, Xiatian Zhu, Zongben Luo, Shuran Liu, Wenyu Li, Jian Yang, and Philip H. S. Torr. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Supplementary Materials

A Metric Definitions

Task 1 classification metrics. For the initial attack failure task, let p_i^{IA} denote the predicted failure probability and $y_i^{\mathrm{IA}} \in \{0, 1\}$ denote the binary label.

The area under the precision–recall curve is

$$\mathrm{AUPRC} = \int_0^1 \mathrm{Precision}(r)\, dr.$$

Precision and recall at threshold τ are

$$\mathrm{Precision}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FP}(\tau)}, \qquad \mathrm{Recall}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FN}(\tau)}.$$

AUROC is

$$\mathrm{AUROC} = \Pr(p_i^+ > p_j^-),$$

where p_i^+ is the predicted risk for a failed initial attack event and p_j^- is the predicted risk for a successful initial attack event.

For threshold-based reporting, the operating threshold is selected on the validation set:

$$\tau^\star = \arg\max_\tau \frac{2\,\mathrm{Precision}_{\mathrm{val}}(\tau)\,\mathrm{Recall}_{\mathrm{val}}(\tau)}{\mathrm{Precision}_{\mathrm{val}}(\tau) + \mathrm{Recall}_{\mathrm{val}}(\tau)}.$$

The predicted binary label is then

$$\hat{y}_i^{\mathrm{IA}} = \mathbf{1}[p_i^{\mathrm{IA}} \geq \tau^\star].$$

Balanced accuracy is

$$\mathrm{BalancedAcc} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\right).$$

For top-k operational triage metrics, let $\mathcal{S}_k$ be the top $k\%$ of test events ranked by predicted failure probability. Then

$$\mathrm{Recall@}k\% = \frac{\sum_{i \in \mathcal{S}_k} y_i^{\mathrm{IA}}}{\sum_{i \in \mathcal{D}_{\mathrm{test}}^{\mathrm{IA}}} y_i^{\mathrm{IA}}}, \qquad \mathrm{Precision@}k\% = \frac{\sum_{i \in \mathcal{S}_k} y_i^{\mathrm{IA}}}{|\mathcal{S}_k|}.$$

Brier score is

$$\mathrm{Brier} = \frac{1}{n}\sum_{i=1}^{n}(p_i^{\mathrm{IA}} - y_i^{\mathrm{IA}})^2.$$

Binary cross-entropy is

$$\mathrm{BCE} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i^{\mathrm{IA}} \log p_i^{\mathrm{IA}} + (1 - y_i^{\mathrm{IA}}) \log(1 - p_i^{\mathrm{IA}})\right].$$

Expected calibration error is

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|,$$

where B_m is the m-th confidence bin, $\mathrm{acc}(B_m)$ is the observed positive frequency in that bin, and $\mathrm{conf}(B_m)$ is the average predicted confidence.

Task 2 regression metrics. For the containment-time task, let h_i denote true containment hours, $\hat{h}_i$ denote predicted containment hours after converting from log space, and $\hat{y}_i^{\mathrm{TTC}}$ denote the predicted log-space target.

Mean absolute error in hours is

$$\mathrm{MAE}_h = \frac{1}{n} \sum_{i=1}^{n} |\hat{h}_i - h_i|.$$

Root mean squared error in hours is

$$\mathrm{RMSE}_h = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{h}_i - h_i)^2}.$$

Median absolute error in hours is

$$\mathrm{MedianAE}_h = \mathrm{median}_i \left(|\hat{h}_i - h_i| \right).$$

Log-space MAE is

$$\mathrm{MAE}_{\log} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i^{\mathrm{TTC}} - y_i^{\mathrm{TTC}}|.$$

Log-space RMSE is

$$\mathrm{RMSE}_{\log} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i^{\mathrm{TTC}} - y_i^{\mathrm{TTC}})^2}.$$

Spearman correlation is

$$\rho_s = \mathrm{corr} \left(\mathrm{rank}(\hat{h}), \mathrm{rank}(h) \right).$$

The coefficient of determination is

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (h_i - \hat{h}_i)^2}{\sum_{i=1}^{n} (h_i - \bar{h})^2}.$$

B Related Work

Initial attack, suppression effectiveness, and suppression difficulty. Initial attack is usually evaluated as a suppression-system outcome, and prior studies show that the definition of "success" depends on containment objective, final-size threshold, response standard, and agency context [50, 85, 15, 47]. Accordingly, operational studies model initial attack with variables that are not always available in public national data, including response time, resource deployment, station capacity, aircraft use, and local suppression objectives [54, 47, 64, 17]. Such work also motivates our access and terrain features, because suppression difficulty is affected by fire behavior, terrain difficulty, roads, response opportunities, and suppression capacity rather than by weather alone [65, 17, 31, 32]. WILDFIREIA differs from these studies by fixing a public event-level contract across the contiguous United States and by excluding operational variables that are not consistently available near discovery time [74, 62, 50].

Containment duration and post-discovery outcomes. Containment time is related to initial attack failure, but it is not the same target: a fire can escape the first response and still be contained

quickly, or it can remain active because of terrain, fuel, weather, resource limits, or reporting conventions [12, 65, 17]. Machine-learning work on containment duration therefore treats the target as a continuous or transformed time variable rather than as a binary escape label [12]. Due to that distinction, WILDFIREIA reports containment time as an auxiliary regression task and keeps the initial attack failure task as the primary ranking problem [14, 40, 50].

Public fire records and environmental covariates. FPA-FOD provides a national event backbone with discovery time, location, cause, final size, and containment fields, while recent FPA-FOD-Attributes work demonstrates the value of attaching physical, biological, social, and administrative covariates to fire occurrence records [74, 62]. Weather and fire danger are commonly derived from gridded meteorological or reanalysis products such as gridMET, ERA5, and MERRA-2, while vegetation, fuel, and terrain features are commonly derived from LANDFIRE and related landscape data [1, 43, 35, 53, 31, 32]. Similarly, access and exposure variables can be attached from vector and demographic products such as OpenStreetMap and WorldPop, but these sources have different spatial resolution, update frequency, and missingness patterns [41, 78, 22, 73]. Thus, our canonicalization pipeline follows the public-data integration direction of prior occurrence and susceptibility work while changing the prediction target from ignition or susceptibility to initial attack outcome [75, 70, 73, 62].

Remote sensing for active fire detection, burned area, and fire mapping. Satellite products are central to modern wildfire monitoring, but they measure thermal activity or burned-area evidence rather than initial attack success [69, 39, 58, 79]. Deep-learning studies and reviews have covered active-fire detection, smoke detection, burned-area segmentation, and satellite-based fire mapping, and these tasks often use image-level or pixel-level labels rather than incident-level suppression outcomes [38, 37, 49, 86]. Because small or newly discovered fires may lack a matched VIIRS detection because of orbital timing, cloud, smoke, geolocation uncertainty, or intensity limits, WILDFIREIA treats discovery-day VIIRS as an optional feature and retains events without matched detections [69, 39, 50]. This choice prevents the benchmark from becoming a satellite-detection benchmark and keeps the sample unit aligned with the fire report [74, 58, 62].

Wildfire spread prediction and fire simulators. Physics-based and coupled fire models such as FARSITE, FlamMap, ForeFire, and WRF-Fire represent fire growth, spread, or fire-atmosphere interaction under explicit environmental assumptions [31, 32, 30, 24]. Data-driven spread benchmarks such as WildfireDB, Next Day Wildfire Spread, WildfireSpreadTS, WSTS+, and Sim2Real-Fire instead evaluate models that predict future fire masks, burned cells, or spread trajectories from remote-sensing and environmental grids [75, 45, 36, 51, 55]. These resources are important for fire-growth modeling, yet their target is future spatial propagation rather than whether an incident remains below an initial attack size objective [63, 3, 5, 4]. Accordingly, WILDFIREIA complements spread benchmarks by using an event-level label, a chronological split, and a forbidden-feature policy built around discovery-time suppression prediction [50, 85, 15].

Models for tabular, temporal, spatial, and spatiotemporal inputs. Initial attack prediction requires comparing models across heterogeneous input forms, because event reports, weather sequences, landscape patches, and patch sequences encode different parts of the same incident [74, 1, 53, 41, 78]. For tabular data, logistic regression, support-vector machines, random forests, gradient boosting, LightGBM, XGBoost, and multilayer perceptrons are common structured-data baselines, and probabilistic evaluation further requires calibration-aware metrics [26, 25, 13, 48, 21, 67, 40, 14, 52]. For temporal data, recurrent, convolutional, and attention-based sequence models provide complementary ways to encode short weather histories, while weather and Earth-system forecasting benchmarks motivate careful handling of spatiotemporal covariates [44, 9, 80, 33, 68, 59, 20]. For spatial and spatiotemporal patches, convolutional encoders, U-Net-style decoders, attention U-Nets, DeepLab-style decoders, transformer segmentation backbones, ConvLSTM/ConvGRU units, PredRNN, TrajGRU, MAU, UTAE, Rainformer, and SwinLSTM-style modules provide established templates for processing gridded or video-like environmental inputs [42, 66, 61, 19, 16, 84, 71, 11, 81, 72, 18, 34, 7, 77]. Our baseline suite adapts these architectures to scalar event-level prediction by replacing dense prediction heads with global pooling and a scalar output, thereby keeping the comparison focused on the same initial attack and containment labels [57, 76, 87, 29, 83].

Table 8: Representative prior work and how WILDFIREIA differs. The table groups related work by target, because the target determines the valid sample unit, split, and leakage constraints.

Line of work	Examples	Typical target	Difference from WILDFIREIA
Initial attack outcome	Korkola et al.; Xu et al.; Butler et al. [50, 85, 15]	IA success or escape	Regional or agency-specific definitions; no shared U.S. multi-modal cache
Suppression difficulty	Lee et al.; Katuwal et al.; Cardil et al. [54, 47, 17]	Response success or difficulty	Often uses operational variables not consistently public nationally
Containment duration	Bhardwaj [12]	Time to containment	Regression target; not a primary IA failure benchmark
Fire occurrence and susceptibility	FPA-FOD-Attributes; WildfireDB; global susceptibility maps [62, 75, 73]	Ignition, occurrence, or susceptibility	Predicts where fires occur, not whether a discovered fire escapes IA
Active fire and burned-area mapping	VIIRS/MODIS/MTBS and remote-sensing DL reviews [69, 39, 79, 38, 37]	Thermal detection or burned pixels	Sensor or pixel labels, not incident-level suppression outcomes
Wildfire spread benchmarks	FireCast, Next Day Wildfire Spread, WildfireSpreadTS, WSTS+, Sim2Real-Fire [63, 45, 36, 51, 55]	Future fire mask or spread trajectory	Forecasts spatial propagation rather than initial attack failure

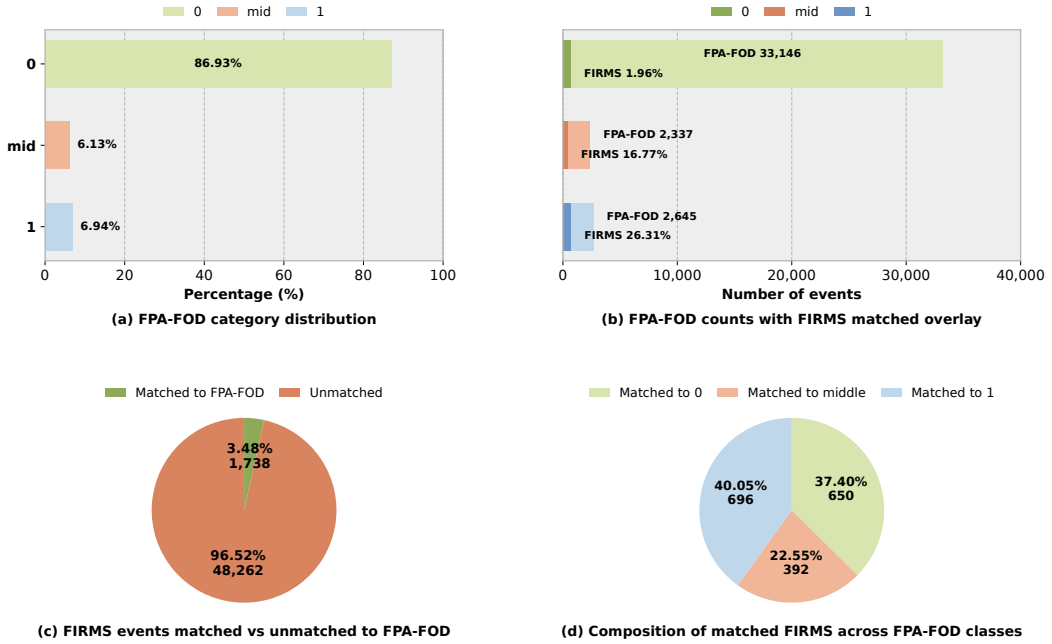

Figure 4: FIRMS/VIIRS matches are a small and biased subset of FPA-FOD natural wildfire events. The full FPA-FOD event set is highly imbalanced, whereas FIRMS-matched events are closer to balanced between success and failure. Thus, FIRMS/VIIRS is used as an input feature source, not as the benchmark sample definition.

C Discussion and Limitations

FPA-FOD events, not FIRMS detections, define the benchmark. A central design choice in WILDFIREIA is that the sample unit is the FPA-FOD natural wildfire event, not the FIRMS/VIIRS active-fire detection. This choice is necessary because the benchmark target is an event-level initial attack outcome: whether a reported wildfire ultimately remains small or escapes initial attack. FIRMS/VIIRS provides valuable discovery-day thermal evidence, but it does not provide a complete event catalog, a final fire size, or a containment outcome. Therefore, we treat FIRMS/VIIRS as an input source attached to FPA-FOD events rather than as the dataset definition.

Figure 4 illustrates why this distinction matters. The FPA-FOD event distribution is highly imbalanced: the small-fire / initial attack success class accounts for most retained natural wildfire events, while escaped fires are rare. However, discovery-day FIRMS/VIIRS matching is much more frequent for larger or more extreme events. Only a small fraction of class-0 events have matched discovery-day FIRMS/VIIRS evidence, whereas the matched fraction is substantially higher for intermediate and escaped-fire events. Thus, restricting the benchmark to FIRMS-matched events changes the event population and makes the label distribution much less rare-event-like. A model trained and evaluated only on this subset can appear much more accurate, but it is solving an easier and less representative problem.

FIRMS-detected subset diagnostic. We additionally define a diagnostic subset to understand the effect of conditioning the benchmark on discovery-day satellite observability. Let

$$z_i^R = \mathbb{1}\{\text{has\_viirs\_detection}_{i,D} = 1\}, \qquad \mathcal{I}_R = \{i : z_i^R = 1\}.$$

Table 9: Prediction-count diagnostic on the FIRMS-detected test subset. The subset contains only FPA-FOD events with matched discovery-day VIIRS detections. Predicted counts are averaged over five seeds.

Model	Test N	True 0	True 1	Pred 0	Pred 1
XGBoost	303	142	161	$78.0_{\pm 7.8}$	$225.0_{\pm 7.8}$
Transformer	303	142	161	$0.2_{\pm 0.4}$	$302.8_{\pm 0.4}$
Swin-UNet	303	142	161	$144.6_{\pm 13.7}$	$158.4_{\pm 13.7}$
SwinLSTM	303	142	161	$105.4_{\pm 63.8}$	$197.6_{\pm 63.8}$

On this subset, we evaluate a FIRMS-focused contract using only FPA-FOD metadata and discovery-day VIIRS features:

$$(M_i, R_i) \to y_i^{\mathrm{IA}}, \qquad i \in \mathcal{I}_R.$$

This diagnostic does not redefine the label: the target remains the FPA-FOD-derived initial attack-failure label y_i^{IA}. The only change is that the evaluated event set is restricted to FPA-FOD events with matched discovery-day FIRMS/VIIRS evidence.

Table 9 shows why this subset is diagnostic rather than primary. The FIRMS-detected test subset contains 303 events, with escaped fires forming the majority class: 161 of 303 events, or 53.1%. This class balance is very different from the full benchmark, where initial attack failure is a rare event. Therefore, FIRMS-only results cannot be directly compared with the full benchmark leaderboard. They answer a conditional question: once a reported wildfire is already visible to VIIRS on the discovery day, how much can the matched thermal signal separate successful initial attack from escaped fires?

This distinction is important for benchmark design. Using FIRMS/VIIRS detection as the event definition would bias the benchmark toward fires that are hotter, larger, more persistent, or easier to observe from orbit. It would exclude many reported FPA-FOD events without matched discovery-day detections, including small or quickly contained fires and events missed because of satellite timing, cloud, smoke, or detection limits. For this reason, WILDFIREIA keeps FPA-FOD as the event backbone and uses FIRMS/VIIRS only as an optional discovery-day input feature. The FIRMS-detected subset is useful for analyzing satellite-observable fires, but it should not replace the full event-level benchmark.

Initial attack failure remains a rare-event prediction problem. The full FPA-FOD-defined benchmark preserves the operational difficulty of initial attack prediction: escaped fires are rare, and many reported fires remain small. This is why we use AUPRC as the primary metric and report calibration metrics in addition to thresholded classification scores. High accuracy on the full benchmark would be easy to obtain by predicting the majority class, but such a model would be useless for identifying fires that require escalation. The benchmark therefore emphasizes rare-failure prioritization rather than aggregate correctness.

Containment duration is a harder auxiliary target. The auxiliary containment-duration analysis shows that the same discovery-time public source stack contains signal about time to containment, but only partially explains the target. The best models achieve modest R^2, around 0.18, indicating that early public variables explain a limited fraction of containment-duration variance. This result is consistent with the nature of the target. Containment duration depends not only on early weather, fuel, topography, access, population context, and satellite fire signal, but also on suppression tactics, crew and aircraft availability, changing weather after discovery, incident command decisions, reporting conventions, and containment criteria. Many of these drivers are not fully observed in our public discovery-time input stack. We therefore interpret containment-duration prediction as an auxiliary stress test rather than as the primary benchmark objective.

Limitations of public event records. FPA-FOD provides a nationwide event backbone, but it is still a reporting-derived dataset. Reported discovery locations may be approximate, final fire size may be recorded with agency-specific conventions, and containment timestamps may be missing or inconsistently reported. Our thresholded initial attack label reduces some ambiguity by separating clearly small fires from clearly escaped fires, but it does not eliminate label noise. Events in the

intermediate 10–50 ha range are marked missing for the binary initial attack target to avoid forcing ambiguous cases into either class.

Limitations of source alignment. All non-FPA-FOD sources are aligned to events through reported discovery date, location, and year. This makes the benchmark reproducible, but it also introduces uncertainty. FIRMS/VIIRS detections depend on overpass timing, cloud and smoke conditions, sensor geometry, and the strictness of the matching radius. gridMET represents coarse weather fields rather than exact fireground conditions. LANDFIRE fuel and vegetation layers are treated as static landscape context, which does not fully capture annual disturbance, treatment, or fuel-moisture dynamics. OSM roads and fire stations are public proxies for access and response proximity, not direct measurements of dispatch decisions, travel time, crew availability, or suppression capacity. WorldPop provides human-settlement context but does not directly encode evacuation priority, asset value, or wildland-urban-interface structure.

Scope of the current benchmark. The current version focuses on natural wildfire events from 2016–2020 and uses a fixed chronological split. This scope reduces ignition-source heterogeneity and supports future-year evaluation, but it does not cover all human-caused wildfire dynamics. Extending the benchmark to all-cause wildfires, adding explicit suppression-resource records, incorporating post-discovery weather evolution, or using high-resolution perimeter growth would enable richer operational questions. However, those additions would also change the information contract. The present benchmark intentionally focuses on what can be attached to a reported natural wildfire near discovery time using reproducible public data.

D Reproducibility Artifacts

The repository is organized around three scripts. `pipeline.py` builds canonical data from raw sources, `dataloader.py` builds tabular, temporal, spatial, and spatiotemporal caches from canonical outputs, and `train.py` trains the baseline models. Experiment directories store the run configuration, model history, metrics, validation and test predictions, and checkpoints for neural models. Summary scripts aggregate metrics by task, representation, model, and seed. This trace is meant to make the benchmark easy to inspect: every reported number should be recoverable from saved `metrics.json` files and prediction parquet files.

E LLM Usage

Large language models were used only for writing assistance, including grammar checking, wording refinement, formatting suggestions, and readability editing. They were not used to generate the core methodology, design experiments, produce experimental results, create evaluation labels, or make routing decisions. All technical claims, mathematical formulations, experimental settings, and reported results were checked and finalized by the authors.