Check for updates

Original Research Article

SCIENCE PROGRESS

Science Progress
Vol. 109(4): 1–12
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00368504261495731
journals.sagepub.com/home/sci
Sage

Early wildfire escalation prediction with temporal validation and statistical benchmarking in a mediterranean region

Gökhan Yüksek[1], Ali Akdagli[2], Serdar Ekinci[3], Davut Izci[4], and Erdal Akin[3,5,6,7]

Abstract

Objective: To evaluate an incident-level framework for predicting whether a newly reported wildfire would reach at least 10 ha, emphasizing temporal validation, statistical benchmarking, and transparent optimizer assessment using only information available at or near incident registration. **Methods:** This retrospective modelling study integrated 1,901 operational wildfire records documented by the Mersin Regional Directorate of Forestry during 2004–2025 with event-day meteorology and Canadian Fire Weather Index variables. Burned area was known for 1,898 incidents, including 68 major fires. Random Forest, Gradient Boosting, and XGBoost were evaluated under default settings, Random Search, Optuna/TPE, and repeated Enhanced Ruppel's Fox Optimization (eRFO), together with regularized logistic regression using the identical early-stage predictors. Three expanding-window folds within 2004–2020 guided development, while 471 incidents from 2021–2025, including 17 major fires, formed a locked recent-year evaluation. **Results:** eRFO produced the highest development PR-AUC for Random Forest (0.142) and Gradient Boosting (0.167), whereas Optuna/TPE achieved the highest XGBoost score (0.170). In the locked recent-year evaluation, the logistic benchmark achieved the highest discrimination (ROC-AUC 0.870; PR-AUC 0.217), compared with ROC-AUC 0.813 and PR-AUC 0.210 for XGBoost–eRFO and ROC-AUC 0.774 and PR-AUC 0.145 for Gradient Boosting–eRFO. Reviewing the highest-ranked 20% of incidents captured 9 of 17 major fires with Gradient Boosting–eRFO and 10 of 17 with XGBoost–eRFO. Out-of-sample SHAP analysis identified wind speed, relative humidity, district, temperature, and drought-related Fire Weather Index components as leading model associations. **Conclusions:** Increased learner and optimizer complexity did not guarantee superior temporal generalization. The principal contribution is a leakage-controlled evaluation framework combining early-available predictors, temporally ordered development, simple-model benchmarking, calibrated probabilities, and interpretable risk ranking for human-in-the-loop prioritization of newly reported wildfires.

Keywords

wildfire initial attack, wildfire escalation, temporal validation, statistical benchmarking, eRFO, Fire Weather Index, rare-event classification

[1]Department of Electrical and Electronics Engineering, Batman University, Batman, Türkiye
[2]Department of Electrical and Electronics Engineering, Mersin University, Mersin, Türkiye
[3]Department of Computer Engineering, Bitlis Eren University, Bitlis, Türkiye
[4]Department of Electrical and Electronic Engineering, Bursa Uludag University, Bursa, Türkiye
[5]Department of Computer Science and Media Technology, Malmö University, Malmö, Sweden
[6]Sustainable Digitalisation Research Centre, Malmö University, Malmö, Sweden
[7]Biofilms Research Center for Biointerfaces (BRCB), Malmö University, Malmö, Sweden

Corresponding authors:
Davut Izci, Department of Electrical and Electronic Engineering, Bursa Uludag University, Bursa 16059, Türkiye.
Email: davutizci@uludag.edu.tr

Erdal Akin, Department of Computer Science and Media Technology, Malmö University, 205 06 Malmö, Sweden.
Email: erdal.akin@mau.se

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Received: 12 May 2026; Revised: 18 August 2026; Accepted: 21 September 2026

1. Introduction

Wildfire research spans ignition forecasting, susceptibility mapping, fire-spread simulation, severity assessment, and operational decision support. These tasks are related but not interchangeable. During initial attack, the practical question is not only where fire may occur, but whether a reported incident is likely to remain small or escape early containment. This distinction is particularly important in Mediterranean environments, where a small proportion of fires accounts for a large share of annual burned area and suppression demand, and where short-term weather, access, response timing, and concurrent fire load influence operational outcomes.[1–4] Incident-level escalation prediction therefore complements the spatial products used for prevention, fuel planning, and preparedness. Most wildfire machine-learning studies nevertheless remain spatial or retrospective. Remote sensing, geographic information systems, terrain, fuels, climate variables, and ensemble learners have been used effectively to identify fire-prone areas and explain broad geographic patterns.[5–13] Such products are valuable for long-term planning, but a gridded hazard score does not directly answer the question faced by dispatch and suppression teams after ignition has been reported: given the information available at that time,[14] which incidents merit the greatest operational attention? A model intended for this purpose must align its predictors with the declared prediction time and avoid quantities that become known only after the fire develops.

International studies of initial attack and containment provide the closest operational context,[15] but their endpoints and predictor sets differ materially from the present formulation.[16] Plucinski modelled Australian grassfire escape using final area ≥100 ha, containment time ≥4 h, or either outcome,[17] with first-situation-report variables that included weather, resource response times, slope, curing state, pasture condition, and estimated fire area at initial attack.[18] Rodrigues et al. modelled the probability of initial-attack containment in Catalonia from georeferenced ignition locations, fire-response variables,[19] and weather conditions,[20] producing spatial containment-probability outputs for first-dispatch planning.[21] Cardil et al. evaluated initial-attack success using simulation-derived Fire Behavior,[22] Terrain Difficulty, and Initial Attack Assessment indices together with suppression response time.[23] The recent Alaska study by Ahajjam et al. combined FIRMS fire detections, 49 geospatial, remote-sensing, and reanalysis features, genetic-algorithm feature selection, and histogram gradient boosting to predict wildfire occurrence, burned area, and duration across multiple forecast horizons.[24] These studies address escape, containment, suppression difficulty, spatial fire behaviour, or broader wildfire forecasting with operational, simulation-based, or remotely sensed inputs. By contrast, the present analysis fixes a single final-size escalation endpoint (≥10 ha) after an incident has been recorded and deliberately restricts predictors to calendar, district, event-day meteorology, and Fire Weather Index information available at or near registration; response time, initial observed fire size, suppression-difficulty indices, and post-event fire behaviour are excluded. The 10 ha threshold is therefore a regional escalation endpoint rather than a universal definition of containment failure. Recent ensemble forecasting and wildfire machine-learning studies also emphasize transparent validation,[25] imbalance control, uncertainty assessment, and interpretability. Hyperparameter configuration remains relevant because tree depth,[26] learning rate, regularization, feature sampling, and estimator count can materially alter rare-event performance. eRFO was developed to improve population diversity and local refinement through adaptive control, logistic-map chaos, opposition-based learning, and Nelder–Mead search.[27] Its usefulness in wildfire modelling should be assessed through learner-specific comparisons with default and standard tuning procedures and, importantly, against a parsimonious statistical benchmark rather than inferred from a single optimized model.

This study evaluates an incident-level framework for predicting whether a reported fire will reach at least 10 ha using calendar, district, event-day meteorological, and Fire Weather Index (FWI) information. The central question is not whether eRFO or an ensemble learner is universally superior, but whether optimized nonlinear learners provide stable recent-year value beyond a regularized logistic benchmark when all models use the same early-available predictors and temporally separated development. The analysis retains 1,898 incidents with known outcomes, uses expanding-window folds for development, and evaluates frozen configurations on 2021–2025 incidents that remain outside preprocessing, optimization, calibration, and threshold selection. Random Forest, Gradient Boosting, and XGBoost are compared under default settings, Random Search, Optuna/TPE, and repeated eRFO optimization; regularized logistic regression provides the statistical benchmark. Evaluation includes discrimination, calibration, uncertainty, threshold behaviour, operational ranking, out-of-sample SHAP interpretation, and geographic summaries.

2. Study Area and Data

2.1. Study area and wildfire archive

The research was conducted within the jurisdiction of the Mersin Regional Directorate of Forestry in south-central Türkiye. The region combines a Mediterranean coastal climate with steep inland terrain, prolonged summer drought, strong seasonal winds, and extensive conifer forests dominated by *Pinus brutia*, *Cedrus libani*, and *Abies cilicica*. These characteristics produce marked variation in ignition frequency and escalation potential across coastal and mountain districts. The operational archive contains 1,901 incident records from 2004–2025, with each row representing a fire documented by suppression personnel. Burned area is known for 1,898 incidents, which together account for approximately 28,900 ha; the median is 0.20 ha, whereas a small number of exceptional fires dominate annual totals. Sixty-eight incidents reached at least 10 ha and 11 reached 500 ha. Figure 1 brings together six complementary views of the archive: annual incident frequency and burned area, geographic coverage, district-level incident counts and escalation rates, monthly FWI distributions, and major-fire rates across FWI categories. The boundary in the temporal plots marks the transition from model development to the independent recent-year evaluation.

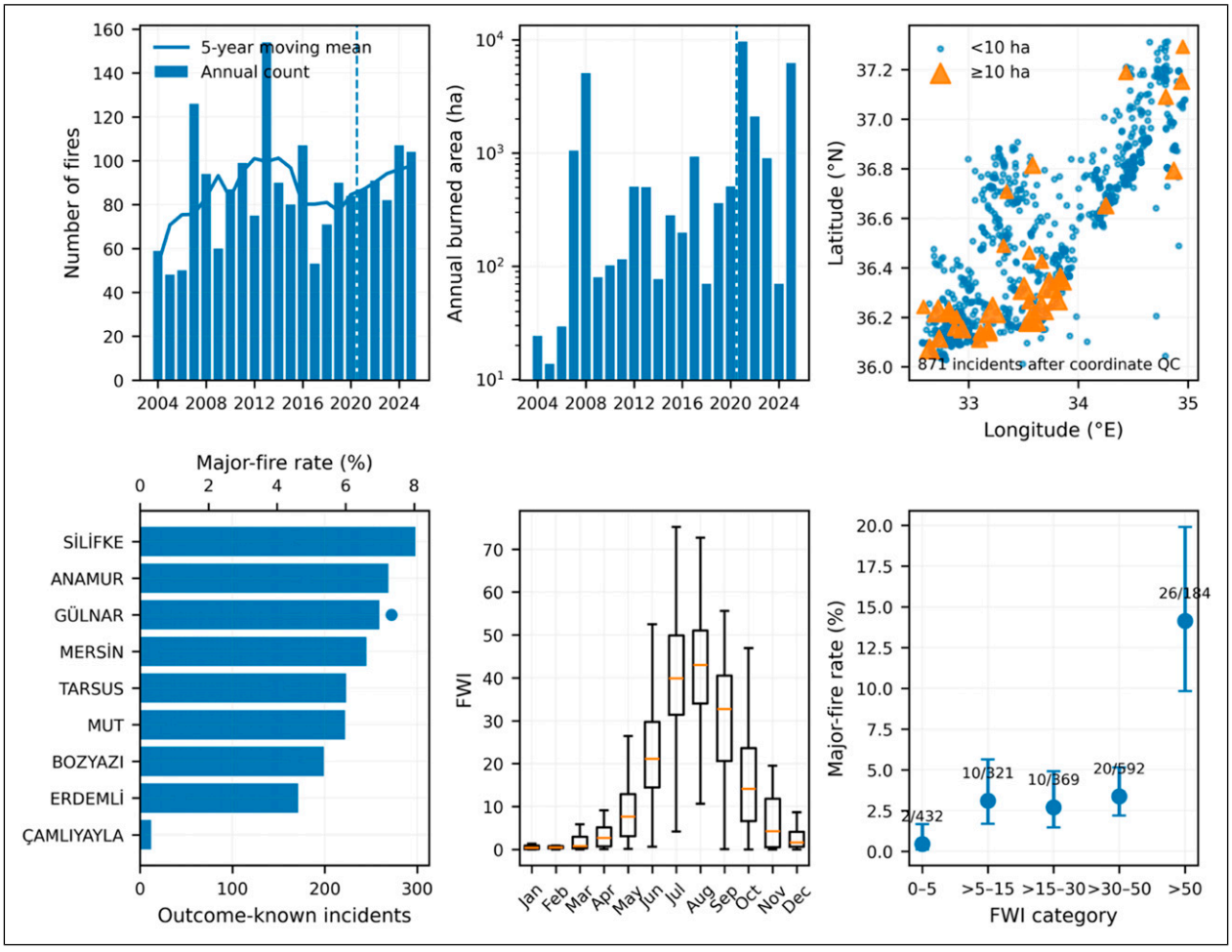


Figure 1. Temporal, geographic, and fire-weather characteristics of the Mersin wildfire archive: upper-left, annual fire counts and 5-year moving mean; upper-middle, annual burned area; upper-right, spatial distribution of the 871 outcome-known incidents retained after coordinate quality control; lower-left, district-level incident counts and major-fire rates; lower-middle, monthly FWI distributions; lower-right, major-fire rates across FWI categories.

2.2. Meteorological data, FWI variables, and endpoint

Daily air temperature, relative humidity, 10 m wind speed, and precipitation were obtained from the NASA POWER archive and matched to incident dates.[28] The daily Canadian FWI sequence was used to derive the Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI), and overall FWI. FFMC characterizes fine-surface-fuel moisture, DMC and DC represent progressively deeper drying, ISI summarizes potential initial spread, and BUI describes the amount of fuel available for combustion. Event-day meteorology and the complete FWI sequence were matched before outcome filtering so that the environmental record remained aligned with the incident archive. The primary endpoint was final burned area of at least 10 ha, yielding 68 positive events or 3.58% of the outcome-known archive. Fires reaching 500 ha were retained only as descriptive context because 11 events do not support stable temporal classification. The predictor set was restricted to information available at incident registration or obtainable from the event-day environmental sequence: day of year, ignition hour, forestry district, temperature, relative humidity, wind speed, precipitation, FFMC, DMC, DC, ISI, BUI, and FWI. Fire duration, response time, first vehicle, cause, and crown-fire information were excluded because they were post-event quantities, incompletely observed, or both. Coordinate values passing the prespecified regional plausibility screen were used solely to describe archive coverage and district heterogeneity, not to estimate ignition susceptibility.

3. Methods

The analytical workflow integrates the incident archive with NASA POWER meteorology and the Canadian FWI sequence, applies training-fold preprocessing, compares three optimized ensemble learners with a regularized logistic benchmark, and evaluates frozen configurations on recent years (Figure 2). All operations capable of influencing model development—including imputation, encoding, hyperparameter optimization, probability calibration, threshold selection, and the SHAP

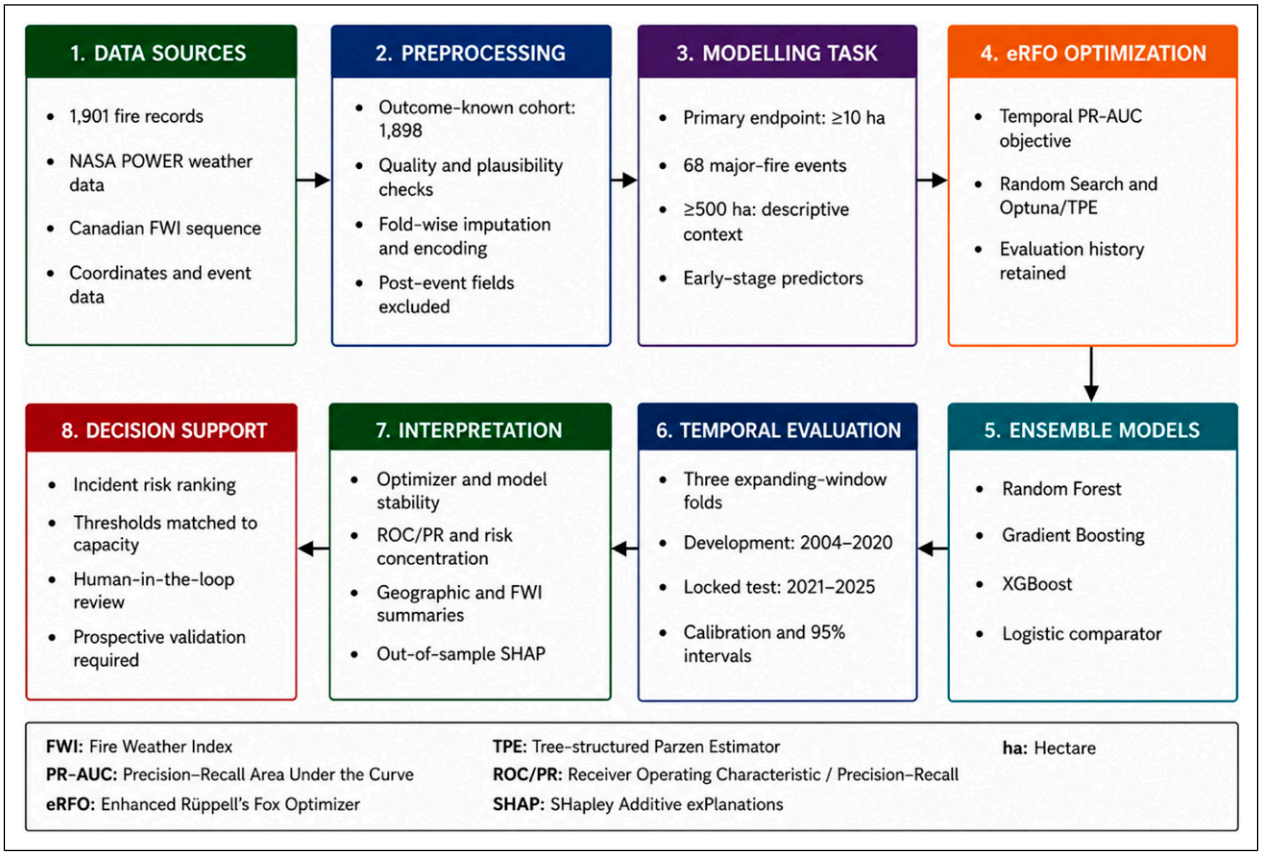


Figure 2. Analytical workflow of the comparative wildfire escalation evaluation framework, comprising data sources, leakage-controlled preprocessing, endpoint definition, optimizer comparison, model-family benchmarking, temporal evaluation, interpretation, and decision-support outputs.

background—were confined to the development period. The eight stages shown in Figure 2 cover data sources, leakage-controlled preprocessing, endpoint definition, optimizer comparison, model-family benchmarking, temporal evaluation, out-of-sample interpretation, and operational decision-support outputs.

3.1. Predictor configuration and preprocessing

Day of year and ignition hour were represented by sine and cosine pairs to avoid artificial discontinuities at year-end and midnight. For day of year d and decimal ignition hour h, the four cyclic terms were calculated according to Equation (1). District was one-hot encoded with previously unseen categories ignored. Numeric imputation used the median estimated from the corresponding training fold, whereas categorical imputation used the most frequent training-fold district. The fitted transformation was then applied unchanged to the associated validation interval and, after final development fitting, to the recent-year evaluation. The outcome-defined approach retained 1,898 of the 1,901 incidents and avoided conditioning the primary analysis on a much smaller subset of operational fields whose missingness varied over time. Weather and FWI predictors were nearly complete, so most incidents required no numeric imputation; nevertheless, the fold-specific procedure ensured that any eligible missing value was handled without leakage. Identical transformed predictors were supplied to the three ensemble learners and the logistic comparator, isolating the effects of learner structure and hyperparameter configuration.

$$s_d = \sin\left(\frac{2\pi d}{365}\right), c_d = \cos\left(\frac{2\pi d}{365}\right),$$
$$s_h = \sin\left(\frac{2\pi h}{24}\right), c_h = \cos\left(\frac{2\pi h}{24}\right). \quad (1)$$

3.2 Temporal development and class imbalance

The 2004–2020 period formed the development sample and contained 1,427 incidents, including 51 major fires. Three expanding-window folds validated on 2011–2013, 2014–2016, and 2017–2020, respectively (Table 1). The locked 2021–2025 period contained 471 incidents and 17 major fires. Its prevalence of 3.61% was close to the development prevalence of 3.57%, while the later calendar position provided a direct test of transportability across changing fire seasons and district composition. Algorithm-level weighting was the primary imbalance strategy during optimizer comparison. Random Forest used balanced subsample weights, Gradient Boosting used balanced observation weights, and XGBoost used the negative-to-positive ratio calculated separately in each training fold. Sensitivity analyses compared no adjustment, class weighting, the Synthetic Minority Oversampling Technique (SMOTE), and SMOTE combined with class weighting.[29] Synthetic observations were generated only after the training partition had been isolated, preventing information from a validation or test interval from entering the resampling process.

Table 1. Archive composition, endpoint frequencies, and temporally ordered development and recent-year evaluation design.

Component	Definition or training period	Training/ total n/positive	Validation/ test period	Validation/ test n/positive	Analytical role
Archive	All recorded incidents, 2004–2025	1,901/68	—	—	Source archive
Outcome cohort	Known burned area, 2004–2025	1,898/68	—	—	Primary modelling cohort
Extreme-fire context	Burned area ≥500 ha	11/—	—	—	Descriptive only
Geographic subset	Coordinate plausibility screen	871/—	—	—	Descriptive mapping
Fold 1	2004–2010	524/14	2011–2013	328/14	Temporal validation
Fold 2	2004–2013	852/28	2014–2016	277/9	Temporal validation
Fold 3	2004–2016	1,129/37	2017–2020	298/14	Temporal validation
Recent-year evaluation	2004–2020	1,427/51	2021–2025	471/17	Locked test

3.3. Learning algorithms and statistical benchmark

Random Forest, Gradient Boosting, and XGBoost were selected because they model nonlinear associations and interactions in tabular data without requiring a prespecified parametric response surface.[30–32] Random Forest reduces variance through bootstrap aggregation and randomized feature selection; Gradient Boosting constructs an additive sequence of trees that corrects residual structure; and XGBoost combines second-order optimization with regularization, row and column subsampling, and explicit controls on child nodes and split gain. The same predictors, temporal folds, and rare-event objective were used for the three ensemble learners. A regularized logistic regression was fitted to the identical early-stage feature set and tuned within the development period to provide an explicit statistical benchmark for model complexity. Because eRFO, Random Search, and Optuna/TPE were used to configure the tree learners, logistic regression did not participate in that optimizer comparison; instead, it was evaluated on the same locked recent-year incidents to test whether nonlinear learner complexity translated into better temporal generalization. This separation allows the optimizer comparison to answer a within-learner tuning question while the logistic benchmark addresses the broader added-value question across model families.

3.4. eRFO-based hyperparameter optimization

eRFO extends the original Rüppell's Fox Optimizer with adaptive search control, logistic-map chaotic diversity, opposition-based learning, and Nelder–Mead local refinement.[27,33] Within the wildfire application, each fox represented a candidate hyperparameter vector and temporal validation performance defined its fitness. The archived executable implementation additionally used elite-preserving pruning and decaying long jumps; its control coefficient followed an iteration-dependent monotone schedule, while the published mechanisms of chaos, opposition, and local refinement were retained. The executable source version, code hash, seed, evaluation history, and elapsed time were fixed for the reported analysis. The optimization criterion was average precision, equivalent to PR-AUC, averaged across the three temporal folds. For a candidate vector x, fold-specific average precision is denoted by $\mathrm{AP_k}(x)$ and the number of folds by K=3. eRFO minimized the loss in Equation (2), so configurations that ranked major fires highly were favoured without allowing the much larger negative class to dominate model selection.

$$J(\mathbf{x}) = 1 - \frac{1}{K}\sum_{k=1}^{K} \mathrm{AP}_k(\mathbf{x}), \qquad K = 3 \tag{2}$$

eRFO was repeated with seeds 7, 42, and 20260719. Random Search and Optuna/TPE used the same learner-specific ranges and temporal folds. Comparator trial budgets matched the median realized eRFO evaluation count within each learner: 13 evaluations for Random Forest and Gradient Boosting and 23 for XGBoost. Bounded Nelder–Mead refinement could increase the evaluation count of an individual eRFO run, so actual counts and elapsed times were retained. The best eRFO configurations used 91 trees and depth 9 for Random Forest; 139 trees, depth 2, and learning rate 0.0553 for Gradient Boosting; and 154 trees, depth 4, and learning rate 0.0155 for XGBoost, together with the sampling and regularization values reported in the computational record.

3.5. Calibration, evaluation, and interpretation

Development out-of-fold probabilities were used for Platt calibration. If $p\mathrm{raw}(x)$ denotes an uncalibrated probability, the calibrated probability in Equation (3) was obtained by applying the logistic function σ to an intercept α and slope β estimated from development predictions only. The same fitted transformation was then applied to recent-year probabilities. An F1-maximizing operating threshold was also selected from development predictions and remained fixed during recent-year evaluation.

$$\begin{aligned} \widehat{p}\,(\mathbf{x}) &= \sigma\left[\alpha + \beta \mathrm{logit}\left\{p_{\mathrm{raw}}(\mathbf{x})\right\}\right], \\ \sigma(z) &= \frac{1}{1+\exp(-z)}. \end{aligned} \tag{3}$$

PR-AUC was the primary rare-event measure and ROC-AUC quantified global discrimination. Brier score summarized probabilistic accuracy, while precision, recall, F1-score, Matthews correlation coefficient, specificity, and the confusion matrix described performance at the selected threshold. Stratified bootstrap resampling with 5,000 repetitions provided 95% confidence intervals for the principal Gradient Boosting–eRFO configuration. Operational usefulness was additionally

assessed by ranking recent-year incidents from highest to lowest predicted probability and calculating the cumulative proportion of major fires captured as the review fraction increased. SHAP values were computed for the development-selected Gradient Boosting model on the 471 recent-year incidents and were interpreted as associations within the fitted prediction function rather than causal effects.[34]

4. Results

4.1. Archive and fire-weather characteristics

Wildfire activity varied substantially across the 22-year record (Figure 1). Incident counts were highest in 2013, while annual burned area was dominated by severe seasons including 2008, 2021, and 2025. The geographic subset followed the coastal and inland structure of the Directorate. Silifke contained the largest number of outcome-known incidents, whereas Gülnar had the highest major-fire rate among districts with a substantial sample, with 19 of 259 incidents reaching at least 10 ha (7.34%). These district contrasts provide context for the location terms in the model but do not constitute out-of-region validation or a susceptibility surface. FWI showed strong seasonality, with the highest median values during July, August, and September. Major fires occurred across the full FWI range, but their frequency increased in the upper categories: 2 of 432 incidents reached 10 ha when FWI was between 0 and 5, compared with 26 of 184 incidents when FWI exceeded 50. The confidence intervals remain wide in the smaller categories, yet the overall gradient supports the inclusion of both immediate weather and accumulated drought indicators rather than reliance on a single annual or seasonal index.

4.2. Model and optimizer performance under temporal evaluation

Development-period optimizer performance and locked recent-year model benchmarking are summarized in Table 2 and Figure 3. For Random Forest, PR-AUC increased from 0.105 with default settings to 0.138 with Random Search, 0.140 with Optuna/TPE, and 0.142 in the strongest eRFO run. Gradient Boosting showed the largest eRFO gain, from 0.124 to 0.167, while XGBoost behaved differently: Optuna/TPE reached 0.170 and the best eRFO run reached 0.137. Across eRFO seeds, variability was small for Random Forest (SD 0.004), moderate for XGBoost (SD 0.011), and larger for Gradient Boosting (SD 0.022). Thus, eRFO was strongest for two learner families during development but did not dominate every learner. The locked comparison additionally included regularized logistic regression so that optimizer gains within the tree learners could be distinguished from model-family generalization across time.

On the locked 2021–2025 period, the regularized logistic benchmark achieved the highest observed discrimination, with ROC-AUC 0.870 and PR-AUC 0.217. Among the ensemble configurations, XGBoost–eRFO produced the highest discrimination (ROC-AUC 0.813; PR-AUC 0.210), followed by Gradient Boosting–eRFO (ROC-AUC 0.774; PR-AUC 0.145). The development-selected Gradient Boosting–eRFO configuration identified 5 of 17 major fires and generated 23 false positives at its fixed threshold, producing precision 0.179, recall 0.294, F1-score 0.222, and specificity 0.949; its bootstrap 95% confidence intervals were 0.672–0.863 for ROC-AUC and 0.074–0.328 for PR-AUC, with 0.065–0.316 for precision and 0.118–0.529 for recall. XGBoost–eRFO captured 8 major fires at its selected threshold, whereas Random

Table 2. Development-period optimizer performance, eRFO seed stability, and locked 2021–2025 model benchmarking.

Learner/configuration	Default	Random Search	Optuna/TPE	eRFO mean ± SD	eRFO best	Seed	Evaluations
Random Forest	0.105	0.138	0.140	0.138 ± 0.004	0.142	20260719	26
Gradient Boosting	0.124	0.144	0.128	0.141 ± 0.022	0.167	42	13
XGBoost	0.121	0.133	0.170	0.129 ± 0.011	0.137	7	13
Locked 2021–2025 evaluation (n=471; 17 major fires)							
Configuration	ROC-AUC	PR-AUC	Brier	Precision	Recall	F1	TP/FP/TN/FN
Logistic comparator	0.870	0.217	0.032	0.148	0.529	0.231	9/52/402/8
Gradient Boosting–eRFO*	0.774	0.145	0.033	0.179	0.294	0.222	5/23/431/12
Random Forest–eRFO	0.755	0.195	0.034	0.150	0.176	0.162	3/17/437/14
XGBoost–eRFO†	0.813	0.210	0.033	0.113	0.471	0.182	8/63/391/9
XGBoost–Optuna/TPE	0.770	0.157	0.033	0.113	0.353	0.171	6/47/407/11

*Development-selected eRFO ensemble configuration. † Learner-specific eRFO sensitivity configuration.

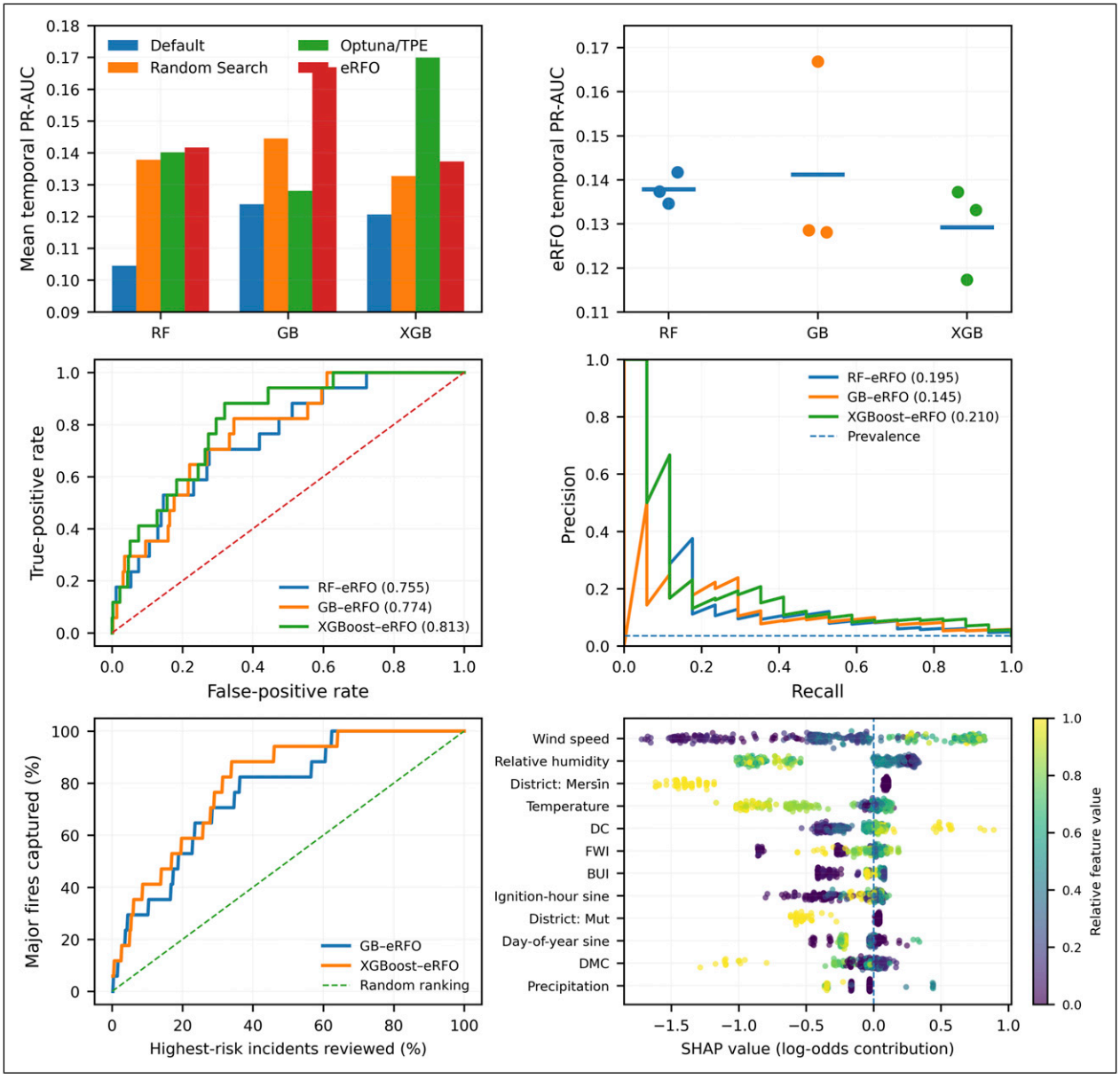


Figure 3. Model optimization and recent-year evaluation: upper-left, development-period optimizer comparison; upper-right, eRFO seed stability; middle-left, ROC curves; middle-right, precision–recall curves; lower-left, major-fire capture as a function of review fraction; lower-right, out-of-sample SHAP distributions for the Gradient Boosting–eRFO model.

Forest–eRFO produced fewer false positives but lower recall. These results show that eRFO found competitive development configurations for selected learners, but development-period optimization did not guarantee superior recent-year discrimination. The logistic result is therefore treated as substantive evidence that learner complexity must earn its place under temporal evaluation, not as proof that logistic regression will be universally superior in other regions or datasets.

4.3 Operational ranking and model interpretation

Risk concentration provided a more direct operational reading than a single binary threshold. Reviewing the highest-ranked 20% of recent-year incidents captured 9 of 17 major fires (52.9%) with Gradient Boosting–eRFO and 10 of 17 (58.8%) with XGBoost–eRFO. Capture increased progressively as the review fraction expanded, indicating that probabilities can support prioritized review even when the appropriate operating point changes with staffing, concurrent incidents, or fire-weather severity. The Brier scores of the evaluated configurations were similar, between 0.032 and 0.034, so the choice among

learners was driven more by ranking and threshold trade-offs than by large differences in aggregate probability error. The SHAP distribution in Figure 3 identified wind speed as the leading transformed predictor, followed by relative humidity, the Mersin district indicator, temperature, DC, FWI, BUI, ignition-hour sine, the Mut district indicator, day-of-year sine, DMC, and precipitation. Higher wind speed generally increased model log-odds, consistent with its influence on spread, flame tilt, spotting, and suppression difficulty. Relative humidity and temperature represented immediate atmospheric controls, while DC and BUI captured deeper seasonal drying. District indicators retained information associated with topography, vegetation, access, and operational context that was not fully represented by the environmental variables. Because these values were calculated for recent-year incidents not used to fit the model, they describe out-of-sample model associations; they do not establish causality.

5. Discussion

The principal comparative finding is that useful early-stage wildfire escalation signal was present in registration-time information, but increased learner and optimizer complexity did not automatically improve temporal generalization. The locked recent-year logistic benchmark achieved ROC-AUC 0.870 and PR-AUC 0.217, exceeding the evaluated ensemble configurations, while XGBoost–eRFO approached the logistic PR-AUC at 0.210. This result changes the interpretation from an optimizer-centred framework to a broader evaluation framework: model development must distinguish gains obtained during hyperparameter search from performance that persists when later years are held completely outside development. Event-day meteorology, FWI components, district, and cyclic time variables still produced useful discrimination, and the out-of-sample SHAP analysis showed that wind speed, relative humidity, temperature, district, and drought-related indices carried substantial signal within the selected Gradient Boosting model. These associations should be read as properties of the fitted prediction function rather than causal effects.

The optimizer comparison remains informative within that broader framing. eRFO identified the strongest development configuration for Random Forest and Gradient Boosting, whereas Optuna/TPE led XGBoost, and the strongest Gradient Boosting eRFO trajectory was not reproduced in every seed. The locked comparison then showed that these within-learner optimization gains did not overcome the simpler logistic decision boundary. In a rare-event setting with 51 development positives and 17 recent-year positives, regularization and lower model variance may be advantageous when nonlinear interactions are unstable across time; however, the present data do not establish a universal preference for linear models. The appropriate lesson is methodological: optimizer performance, learner complexity, calibration, ranking stability, interpretability, and the consequences of false-positive and false-negative decisions should be evaluated separately. Gradient Boosting–eRFO maintained high specificity and a relatively low false-positive count, whereas XGBoost–eRFO captured more major fires but flagged more non-major incidents. Those operating characteristics may matter differently under limited review capacity, severe fire weather, or changing resource availability even when a parsimonious model has the highest overall discrimination.

5.1. Operational implications

The most direct operational use is a ranked list of newly reported incidents requiring closer review during periods of high demand. The model does not replace dispatch protocols or incident-command judgement; it adds a consistent environmental and temporal signal that can be combined with locally observed fire behaviour, access, resource availability, and concurrent workload. District and FWI summaries also offer planning value by identifying parts of the archive and fire-weather ranges in which escalation occurred more frequently. Operational implementation would require timestamped local observations or forecasts, a clearly defined prediction moment, predeclared review rules, and prospective monitoring of calibration and false-alarm burden.

5.2. Limitations and future research

The study represents one Mediterranean forestry region and includes 68 major fires, with 17 positives in the recent-year evaluation. These counts permit temporal analysis but yield wide precision–recall intervals and limit fine-grained subgroup inference. NASA POWER provides a consistent daily environmental sequence, yet a retrospective daily value cannot reproduce every feature of a live forecast or local station observation. Coordinates were available for part of the archive, and district indicators may combine environmental, topographic, access, reporting, and operational differences that cannot be separated with the available fields. Potentially informative variables such as suppression resources, access constraints, and early observed fire behaviour require reliable timestamps and more complete recording before they can be incorporated into a genuine early-stage model. Future research should test the framework prospectively and across additional forest regions.

Multi-region analysis would increase the number of major fires, permit leave-one-region-out validation, and show whether the selected eRFO configurations remain transferable. Higher-resolution meteorological observations, forecast FWI, fuel maps, topographic descriptors, and time-stamped satellite or thermal indicators may improve discrimination when their acquisition time matches the prediction point. Operational studies should also compare alternative escalation thresholds and cost-sensitive decision rules with fire managers, because the acceptable balance between false alarms and missed events depends on concurrent fire load, available resources, and the consequences of delayed response.

6. Conclusions

This study evaluated early wildfire escalation prediction using a leakage-controlled, temporally ordered framework rather than treating optimizer performance as sufficient evidence of model superiority. The outcome-known cohort included 1,898 fires, the endpoint was fixed at ≥10 ha, and predictors determined after fire development were excluded. During development, eRFO produced the strongest configurations for Random Forest and Gradient Boosting, whereas Optuna/TPE led XGBoost. On the locked 2021–2025 evaluation, however, the regularized logistic benchmark achieved the highest observed discrimination (ROC-AUC 0.870; PR-AUC 0.217), while XGBoost–eRFO reached ROC-AUC 0.813 and PR-AUC 0.210 and Gradient Boosting–eRFO reached ROC-AUC 0.774 and PR-AUC 0.145. The findings therefore show that hyperparameter optimization and nonlinear complexity can improve development performance without guaranteeing superior temporal transportability. The practical contribution is the evaluation structure itself: a clearly defined incident-level endpoint, early-available predictors, explicit simple-model benchmarking, transparent optimizer assessment, calibrated probabilities, risk-concentration analysis, and out-of-sample interpretation. The framework is intended for human-in-the-loop prioritization rather than automated suppression decisions; prospective testing and external regional validation remain necessary before operational deployment.

Ethical considerations

Ethical approval was not required because the study used historical institutional wildfire incident records and involved no human participants, human tissue, personal health data, or animals.

Acknowledgements

The authors gratefully acknowledge the Mersin Regional Directorate of Forestry for providing the wildfire incident records and granting permission for their anonymized public release.

ORCID iDs

Serdar Ekinci https://orcid.org/0000-0002-7673-2553
Davut Izci https://orcid.org/0000-0001-8359-0875
Erdal Akin https://orcid.org/0000-0002-2223-3927

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research received no specific grant from any funding agency in the commercial, or not-for-profit sectors. The computational hardware infrastructure used for the simulations in this study was funded by Mersin University BAP Unit under Grant Contract No: 2024-TP3-5172.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The Mersin Regional Directorate of Forestry formally authorized public sharing of the anonymized incident records and coordinates used in this study. The complete reproducibility archive, comprising analysis code, the executable eRFO implementation, locked temporal protocols, the anonymized modelling table, prediction probabilities, result tables, and code-generated figures, is deposited in Zenodo under DOI: https://doi.org/10.5281/zenodo.21455927. The software is released under the MIT License and the anonymized data under the Creative Commons Attribution 4.0 International License (CC BY 4.0).[35]

References

1. Alcasena F and Vega-García C. Predicting wildfire losses is crucial for confronting the European wildfire crisis. *2023 8th International Conference on Smart and Sustainable Technologies*. IEEE, 2023, pp. 1–5.
2. Cullen AC, Goldgeier BR, Belval E, et al. Characterising ignition precursors associated with high levels of deployment of wildland fire personnel. *Int J Wildland Fire* 2024; 33. https://doi.org/10.1071/WF23182
3. Ortega M, Silva FR and Molina JR. Modeling fuel break effectiveness in southern Spain wildfires. *Fire Ecol* 2024; 20: 40. https://doi.org/10.1186/s42408-024-00270-2
4. Alawode GL, Gelabert PJ and Rodrigues M. A spatially explicit containment modelling approach for escaped wildfires in a Mediterranean climate using machine learning. *Geomat Nat Haz Risk* 2025; 16: 2447514. https://doi.org/10.1080/19475705.2024.2447514
5. Abujayyab SKM, Kassem MM, Khan AA, et al. Wildfire susceptibility mapping using five boosting machine-learning algorithms: the Mediterranean region of Turkey. *Adv Civ Eng* 2022; 2022: 3959150. https://doi.org/10.1155/2022/3959150
6. Aksoy E, Kocer A, Yilmaz İ, et al. Assessing fire risk in wildland–urban interface regions using a machine-learning method and GIS data. *Istanbul's European side. Fire* 2023; 6: 408.
7. Novo A, Dutal H and Eskandari S. Fire susceptibility modeling and mapping in Mediterranean forests of Turkey: fuel, climatic, topographic, and anthropogenic factors. *EuroMediterr J Environ Integr* 2024; 9: 655–679. https://doi.org/10.1007/s41207-024-00475-6
8. Agrawal N, Nelson PV and Low RD. A novel approach for predicting large wildfires using machine learning, remote sensing and atmospheric reanalysis data across the United States. *Remote Sens* 2023; 15: 5501. https://doi.org/10.3390/rs15235501
9. Alkan AH, Akinci H and Zeybek M. Comparison of diverse machine-learning algorithms for forest-fire susceptibility mapping in Antalya, Türkiye. *Adv Space Res* 2024; 74: 647–667. https://doi.org/10.1016/j.asr.2024.04.018
10. Göksu Bektaş A, Karaş İR, Akay AE, et al. Using machine-learning algorithms to predict forest-fire probability in the Mediterranean region of Türkiye. *Forestist* 2025; 75: 1–14. https://doi.org/10.5152/forestist.2024.24022
11. Eker R and Aydın A. Predicting potential fire severity in Türkiye's diverse forested areas: a SHAP-integrated random-forest classification approach. *Stoch Environ Res Risk Assess* 2024; 38: 4607–4628. https://doi.org/10.1007/s00477-024-02820-1
12. Iban MC and Sekertekin A. Machine-learning-based wildfire susceptibility mapping using remotely sensed fire data and GIS: Adana and Mersin provinces, Turkey. *Ecol Inform* 2022; 69: 101647. https://doi.org/10.1016/j.ecoinf.2022.101647
13. Sari F. Identifying anthropogenic and natural causes of wildfires by maximum-entropy-method-based ignition susceptibility distribution models. *J For Res* 2023; 34: 355–371. https://doi.org/10.1007/s11676-022-01502-4
14. Arienti MC, Cumming SG and Boutin S. Empirical models of forest-fire initial-attack success probabilities: effects of fuels, linear features, fire weather, and management. *Can J For Res* 2006; 36: 3155–3166. https://doi.org/10.1139/x06-188
15. Morin AA, Albert-Green A, Woolford DG, et al. The use of survival-analysis methods to model the control time of forest fires in Ontario, Canada. *Int J Wildland Fire* 2015; 24: 964–973. https://doi.org/10.1071/wf14158
16. Paudel A, Martell DL and Woolford DG. Factors affecting the timing of initial-attack resource dispatch to forest fires in northeastern Ontario. *Int J Wildland Fire* 2019; 28: 15–24. https://doi.org/10.1071/wf18058
17. Collins KM, Price OF and Penman TD. Suppression-resource decisions are the dominant influence on containment of Australian forest and grass fires. *J Environ Manage* 2018; 228: 373–382. https://doi.org/10.1016/j.jenvman.2018.09.031
18. Plucinski MP. Modelling the probability of Australian grassfires escaping initial attack to aid deployment decisions. *Int J Wildland Fire* 2013; 22: 459–468. https://doi.org/10.1071/wf12019
19. Tremblay P-O, Duchesne T and Cumming SG. Survival analysis and classification methods for forest-fire size. *PLoS One* 2018; 13: e0189860. https://doi.org/10.1371/journal.pone.0189860
20. Korkola K, Wheatley M, Beverly J, et al. A comparative analysis of wildfire initial-attack containment objectives and modelling strategies in Ontario, Canada. *Int J Wildland Fire* 2024; 33. https://doi.org/10.1071/WF24104
21. Rodrigues M, Alcasena F and Vega-García C. Modeling initial-attack success of wildfire suppression in Catalonia, Spain. *Sci Total Environ* 2019; 666: 915–927. https://doi.org/10.1016/j.scitotenv.2019.02.323
22. Kim SY, Jun C and Na W. Ensemble-based forecasting of wildfire potentials using relative index in Gangwon Province, South Korea. *Ecol Inform* 2025; 86: 103021. https://doi.org/10.1016/j.ecoinf.2025.103021
23. Cardil A, Jiménez-Ruano A, Monedero S, et al. Assessing the suppression difficulty of wildland fires for initial-attack response. *Int J Wildland Fire* 2025; 34. https://doi.org/10.1071/WF24160
24. Ahajjam A, Allgaier M, Chance R, et al. Enhancing prediction of wildfire occurrence and behavior in Alaska using spatio-temporal clustering and ensemble machine learning. *Ecol Inform* 2025; 85: 102963. https://doi.org/10.1016/j.ecoinf.2024.102963
25. Xu Z, Li J, Cheng S, et al. Deep learning for wildfire risk prediction: integrating remote sensing and environmental data. *ISPRS J Photogramm Remote Sens* 2025; 227: 632–677. https://doi.org/10.1016/j.isprsjprs.2025.06.002

26. Ejaz N and Choudhury S. A comprehensive survey of the machine-learning pipeline for wildfire risk prediction and assessment. *Ecol Inform* 2025; 90: 103325. https://doi.org/10.1016/j.ecoinf.2025.103325
27. Yüksek G. Enhanced Ruppel's Fox Optimizer (eRFO): an improved metaheuristic with adaptive control and chaotic diversity. *Electrica* 2026; 26: 0298. https://doi.org/10.5152/electrica.2026.25298
28. NASA POWER Project. Prediction of Worldwide Energy Resources. https://power.larc.nasa.gov (2026), accessed 19 July 2026.
29. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002; 16: 321–357. https://doi.org/10.1613/jair.953
30. Chen T and Guestrin C. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. : ACM, 2016, pp. 785–794.
31. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32. https://doi.org/10.1023/a:1010933404324
32. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001; 29: 1189–1232. https://doi.org/10.1214/aos/1013203451
33. Braik M and Al-Hiary H. Rüppell's fox optimizer: a novel meta-heuristic approach for solving global optimization problems. *Cluster Comput* 2025; 28: 292. https://doi.org/10.1007/s10586-024-04950-1
34. Lundberg SM and Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017; 30: 4765–4774.
35. Yüksek G, Akdagli A, Ekinci S, Izci D and Akin E. *Reproducible eRFO-Assisted Wildfire Escalation Analysis (Version 1.0.0) [Computer software]*. Zenodo, 2026. https://doi.org/10.5281/zenodo.21455927.