Localized Forest Fire Risk Prediction: A Department-Aware Approach for Operational Decision Support

Nicolas Caron[1], Hassan Noura[1], Christophe Guyeux[1], and Benjamin Aynes[2]

[1]Université Marie et Louis Pasteur, CNRS, institut FEMTO-ST, F-90000 Belfort, France, France
[2]SAD Marketing, Lille, France

Abstract

Forest fire prediction involves estimating the likelihood of fire ignition or related risk levels in a specific area over a defined time period. With climate change intensifying fire behavior and frequency, accurate prediction has become one of the most pressing challenges in Artificial Intelligence (AI). Traditionally, fire ignition is approached as a binary classification task in the literature. However, this formulation oversimplifies the problem, especially from the perspective of end-users such as firefighters. In general, as is the case in France, firefighting units are organized by department, each with its own terrain, climate conditions, and historical experience with fire events. Consequently, fire risk should be modeled in a way that is sensitive to local conditions and does not assume uniform risk across all regions. This knowledge is generally ignore in the current researches on wildfire prediction with AI. This paper proposes a new approach and a new database that tailors fire risk assessment to departmental contexts, offering more actionable and region-specific predictions for operational use. With this, we present the first national-scale AI benchmark for metropolitan France using state-of-the-art AI models on a relatively unexplored dataset. Finally, we offer a summary of important future works that should be taken into account. Supplementary materials are available on github *NicolasCaronPro/Localized-Forest-Fire-Risk-Prediction-ADepartment-Aware-Approach-for-Operational-DecisionSupport*.

1 Introduction

The immediate cost of wildfire suppression, compensation for damaged property, and the aftermath of lost agricultural productivity and tourism put a strain on local economies. For instance, the 2017 wildfires in the Mediterranean region resulted in estimated economic damages exceeding €200 million, according to *Le Monde (2017)*. From a human perspective, wildfires in France often lead to evacuations, sometimes involving thousands of residents and tourists. Health risks related to smoke inhalation and post-fire water contamination are also significant concerns [1, 2, 3]. Ecologically, these fires can cause irreversible harm to unique ecosystems. Additionally, wildfires contribute to topsoil erosion, threatening agriculture and leading to landslides in hilly terrains. Lastly, the emissions from these fires contribute not only to local air pollution but also to France's national carbon footprint, further complicating efforts to combat climate change. According to *Reporterre (2024)*, more than 4,000 hectares have burned in 2024.

1.1 Problem formulation

In this work, the problem of predicting the future forest fire risk is considered. We define the wildfire prediction in definition 1.

Definition 1 ***Wildfire prediction*** *: Predict the future risk of wildfire ignition (or a linked-value) V in a particular area A, for a particular time range T. Knowing a particular set of characteristics F, we can mathematically define the risk of wildfire as:*

$$V(A, T) = R(F(A, T))$$

where R is a particular set of functions or an algorithm that calculates the risk value.

1.2 State of the art

In the literature, we found publicly available datasets used to address the wildfire prediction problem.

Kondylatos et al. [4] released a dataset for wildfire hazard prediction in Greece, covering the period 2009–2021 at a daily spatial resolution of approximately $1, \mathrm{km}^2$. The dataset integrates meteorological variables, land cover, and population data. Each file is about 23 GB in size, which raises significant challenges for memory usage and accessibility.

The SeaFire Cube platform [5] offers 21 years of global data (2001–2021), 8-day intervals, $0.25°$ resolution. It includes atmospheric, climate, vegetation, socioeconomic, and

fire variables (burned area, CO_2). Easy to access with tutorials, it suits seasonal prediction but not fine-scale management. Michail et al. [6] used it with GraphCast [7], reaching global AUPRC 0.64, but only 0.20 in Europe, highlighting poor regional generalization and the need for domain-specific models.

The EO4WildFires benchmark [8] targets severity forecasting (affected area). It combines multi-sensor series (Sentinel-1, Sentinel-2, NASA Power) for 31,730 events across 45 countries (2018–2022), annotated with EFFIS ($\sim$25 GB in one file). It is rich in features (temperature, precipitation, soil moisture, snow), but large volume hampers usability.

Mesogeos [9] provides a dataset for the Mediterranean region spanning 2006–2022, with a daily spatial resolution of approximately $1, \mathrm{km}^2$ and 27 variables covering meteorology, vegetation, land cover, and human activity. It includes records of ignitions and burned areas larger than 30 ha. Despite the availability of extraction tools and a leaderboard, the dataset's large volume poses significant challenges. Moreover, excluding smaller fires also removes important signals, such as detection time and response efficiency. Among the evaluated models (LSTM, GTN, and Transformer), LSTM achieved the best performance.

Across all these datasets, the wildfire risk prediction problem is typically treated as a binary classification task (using the predicted probability to generate risk maps). Employing a very fine spatial resolution (1 km $\times$ 1 km) allows for a detailed representation of spatial variables (e.g., land cover, vegetation dryness). However, this also introduces a strong bias in the evaluation of metrics, since only a tiny fraction of the test samples correspond to fire events (e.g., only 1,228 fire pixels in [4] out of several million pixels). Furthermore, the aspect of model calibration has rarely been studied and models often replicate known high-risk zones, while it has been proved that accuracy drops at fine scales [10, 11]. Another limitation is the large memory consumption required, which restricts their use in practice. Finally, one element not provided by current datasets is information on the organization of firefighting units within each country, even though risk prediction should ideally be aligned with this scale of organization.

1.3 Contribution

Our contribution rests on two main points:

1. Unlike existing public datasets, ours integrates the administrative structure of French departments, which aligns with the fire service organization and makes predictions more realistic for operational use. We incorporate calendar features (day, holidays, etc.), detailed forest cover information (tree species), refined land cover, and a comprehensive set of fire indices (FWI, Nesterov, Angström). Spatial data are provided in two formats to facilitate categorical handling. All features are aggregated into an $2, \mathrm{km} \times 2, \mathrm{km} \times 1$ day xarray. To manage memory, only one departmental cube is loaded at a time—this increases aggregation time but enables flexible study areas ranging from a single department to the whole of France. We demonstrate usage through fire count prediction with an ordinal classification scheme, though the dataset also supports broader applications in risk management and analysis. Similar studies to [4, 9] can likewise be conducted.

2. Binary wildfire danger ignores France's regional heterogeneity: two ignitions in low-risk Brittany may indicate a crisis, while the same in Mediterranean areas is routine. Standard models flatten these contrasts, trivializing near-certain summer fires in the south and exaggerating sparse events elsewhere. This uniformity misguides resource allocation. The prediction of the number of fires (or of the total burned area) is very random (4 vs. 5 vs. 6 fires), which may also limit the model's convergence. We thus propose the first national AI benchmark for metropolitan France, replacing binary labels with a region-aware multi-class scheme.

2 Constructing the database

In this section, we present the original fire source we used to construct the database along with the original features sources.

2.1 Processing Fire occurrence and Burned Area

BDIFF is an internet application designed to centralize all data on forest fires across French territory since 2006 and make this information available to the public and state services. Fire locations are referenced by the name of the city where they occurred; coordinates are not provided. Fire are defined, notably, by the BA.

Data from June 12, 2017, to December 31, 2024 were collected. The years 2017–2020 and 2022 were used for training, 2021 and 2024 for validation, and 2023 for testing. Due to the severity of fire activity in 2022, it was included in the training set, while a more typical years was selected for validation. We did not include 2024 in the test split but used it for validation, because it was a relatively low-fire year (about 1,600 fires vs. 2,300 in 2023), which would make it a less reliable test year.

City names were used to geolocate fire points and to generate $2, \mathrm{km} \times 2, \mathrm{km}$ rasters for each department. Two daily targets were defined: Fire Occurrence (FO) and Burned Area (BA). Instead of treating these as binary or regression tasks,

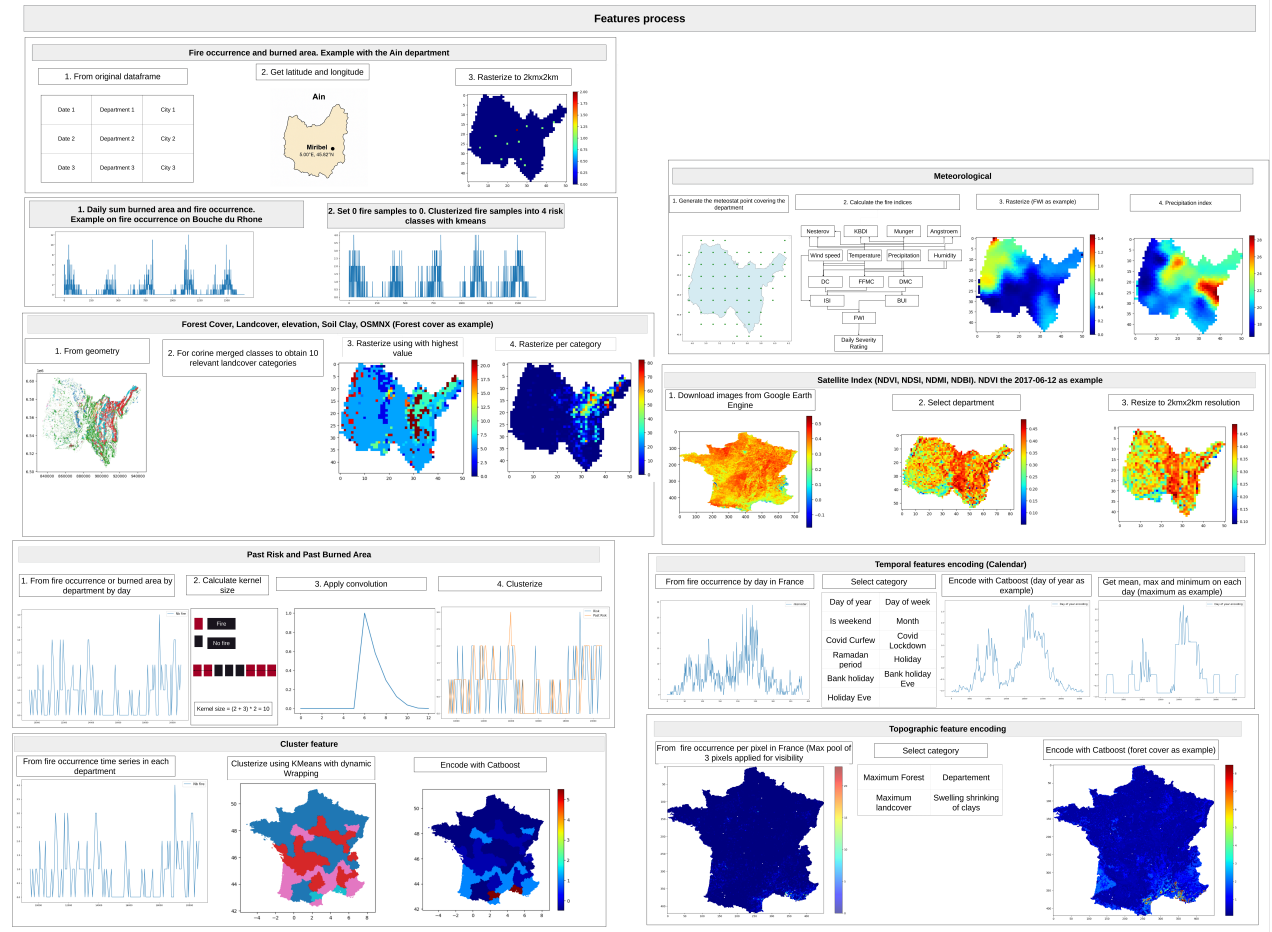


Figure 1: Database construction process applied in this study. Apart from the target-related process, for which we show Bouches-du-Rhône, the department shown is Ain.

we reformulated them as multi-class problems by constructing an ordinal five-class signal for both FO and BA using K-Means clustering. Class 0 corresponds to the absence of fire, while positive samples were grouped into four levels (Normal, Medium, High, Extreme). This approach emphasizes typical rather than absolute values, thereby adapting to departmental variability and improving interpretability for firefighting operations. All predictions are made at the daily scale.

Figure 2 compares class distributions between the Mediterranean basin and the rest of France. Daily FO levels are similar nationwide (except for class 3 and 4, which are concentrated in the Mediterranean). For the BA, class 4 shows high variance, suggesting that extreme events are scattered and department-specific. This supports the idea that our clustering method enhances predictability but may obscure moderate events when extreme cases dominate a department's historical profile. The resulting class distributions show a strong imbalance, with class 0 dominating. Class 1 is the most frequent among positive classes for both targets. Superior classes (2,3 and 4) are underrepresented, specially in BA.

2.2 Processing Features

The features used in this database fall into four categories (Table 1): *Meteorological*, *Topographic*, *Socio-Economic*, and *Historical*. Most were rasterized into 2 km resolution datacubes for consistency, except for historical data, which were computed at the site level (see Figure 1).

- Meteorological data were calculated on an 11×11 grid per department. Fire indices followed standard methods and were reset annually. The precipitation index, based on Chen et al. [12], captures short-term rainfall variation and was

Table 1: Summary of features used in this study. '-' means the same as above.

Variables	Frequency	Source	Variables	Frequency	Source
Meteorological			**Topographic**		
Temperature	12h, 16h	Meteostat	Elevation	Static	CourbesDeNiveau (IGN)
Dew Point	-	-	Forest landcover	-	BDForet (IGN)
Precipitation	-	-	Landcover	-	Corine
Wind Direction	-	-	NDVI, NDSI, NDMI, NDBI, NDWI	15 days	GEE (landsat 1+2)
Wind Speed	-	-	Swelling-shrinking of clays	-	-
Precipitation in Last 24 hours	-	-			
Snow height	-	-			
Sum of last 7 days rain drop	-	-			
Day since last rain	12h	-			
Nesterov	-	firedanger			
Munger	-	-			
KBDI	-	-			
Angstroem	-	-			
BUI, ISI, FFMC, DMC, FWI,	-	-			
Daily severity rating	-	-			
Precipitation Index last 3, 5, and 9 days	-	Calculated			
Socio-Economic			**Historical**		
Highway	Static	BDRoute (IGN)	Past risk	Daily	Calculated
Population	-	Kontur	Past risk BA	-	-
Calendar	Daily		Cluster	Static	-
			Department	-	-

computed directly on the 3D raster.

- Past risk and **BA** features represent prior fire activity, processed using cubic kernel convolution (5 classes) and shifted by one day. The kernel size was based on the average fire sequence duration, with a 3-day inactivity threshold defining sequence boundaries.

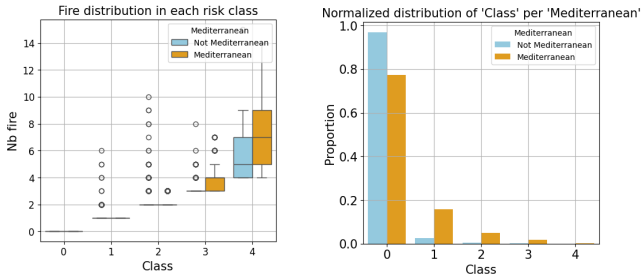
- Cluster variables were derived using K-Means clustering with Dynamic Time Warping [13], grouping departments with similar fire patterns.

- Calendar features capture daily context (e.g., weekdays, holidays, curfews) and were encoded using CatBoost [14] based on departmental fire totals. Aggregated statistics (mean, sum, min, max) were computed across encoders to represent calendar risk.

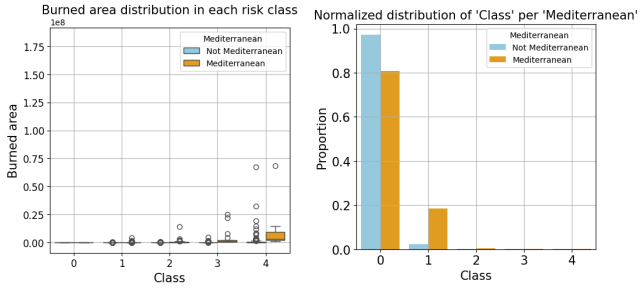
- Categorical features were also CatBoost-encoded using fire counts per pixel. Where needed, rasters were also split by subcategories (e.g., primary vs. secondary roads) to preserve detail lost at 2 km resolution. NDVI, NDSI, NDBI, NDMI, and NDWI (Normalized Difference Vegetation, Snow, Build-up, Moisture, and Water Index) are satellite indices calculated using Landsat 1 and 2 bands. These indices represent key environmental characteristics such as vegetation density, snow cover, built-up areas, soil moisture, and surface water presence—all of which are relevant for assessing fire risk. The Corine database originally contains 44 land cover classes, but some classes were merged in order to aggregate similar information (such as snow and rock simulate no vegetation area). Therefore, logical reduction was applied to retain only the land cover classes most relevant to forest fires: Urban, transport, forest, natural vegetation, agriculture, grassland, natural non-vegetation area, littoral, water area, and wetland. Code is available in the supplementary materials.

Except for **Department**, **Historical**, and **Calendar** data, all spatial features were aggregated (min, max, mean) by cluster. Finally, all variables—aside from *Past risk* and *Past BA*—were standardized.

Storing all x-array files requires about 150 GB, with each data cube averaging 1.5 GB. Building all x-arrays can take several days - up to a week. We used a Dell Precision 7780 with 32 GB of RAM and a 13th Gen Intel® Core™ i7-13850HX (28 cores). We do not currently have the rights to publish the xarray files, but the code for downloading and reconstructing them is released on GitHub. Automated downloading of raw data is provided where possible; however, some features (e.g., CORINE and GEE) require authentication. All steps is documented.



(a)



(b)

Figure 2: Distribution of FO (a) and BA (b) classes risk in the Mediterranean basin and the rest of France. Histograms have been computed relative to the category -Mediterranean (orange) or not (blue)- taking all horizons together.

2.3 Features selection

This process produced a total of 261 features. Those with no variance and those highly correlated with each other (Pearson, Spearman, and Kendall correlation threshold of 0.95) were removed, retaining those with the highest variance. This selection resulted in 162 features. The final feature list is accessible in the supplementary materials. The final dataset is composed of 2,758 unique dates (with 2,192 dates corresponding to a fire in France), across 92 departments (Paris was not include due to lack reliability concerning fire. Corse was not include due to the different vegetation compared to the rest of France), for a total of 253,736 rows.

3 Models

The models used are described in detail in table 2, description of the deep learning layer parameters, and tree-based configuration is available in the supplementary materials. The data imbalance problem was addressed by testing various proportions of class 0, starting from 5% to 100%, with a step of 5%. The percentage selected maximize the Intersection over Union on the validation set. Time series models (LSTM, GRU, DilatedCNN) were tested with 10 days sequences. All models were trained with CrossEntropy loss. The learning rate for deep learning models was set to 0.00005. All model parameters were tuned based on validation performance, mainly on horizon 0. The FWI system is based on the classification proposed by Dupire et. al [15]. Standard deviation has been calculating at each step of the training process with 5 runs.

Table 2: Models used for the benchmark.

Model	**Description**
Logistic Regression	A linear model that estimates class probabilities using the logistic function, optimized via cross-entropy. Simple and interpretable but limited for highly non-linear relationships.
XGBoost	An optimized gradient boosting algorithm using second-order derivatives and regularization for efficiency and robustness. Fast and accurate, though often requires careful hyperparameter tuning.
CatBoost	A gradient boosting method on decision trees that handles categorical features natively with ordered encodings. Strong performance on tabular data with minimal tuning.
MLP	A Multi-Layer Perceptron applying three ReLU-activated fully connected layers, outputting class probabilities through a Softmax head.
DilatedCNN (1D)	Stacks dilated 1D convolutions with normalization and dropout before a linear classification head producing class scores and Softmax probabilities from the final time step.
GraphCastGRU	Encodes temporal signals with a GRU, feeds embeddings into `GraphCast`, and applies linear layers plus Softmax activation for graph-level class predictions.
LSTM	Processes sequences with an LSTM block (optionally layer-normalized), applies dropout and ReLU-activated linear layers, and outputs class probabilities via a Softmax classifier.
GRU	Uses a stacked GRU to summarize temporal dynamics, applies normalization and dropout, then maps the hidden state through linear layers followed by Softmax for classification outputs.

We evaluate the models' performance on each target using two main metrics:

1. Binary F1 score, which measures the performance of predicting the presence of at least one fire. Additionally, we report the precision and recall scores.

2. Silva et al. [16] introduce AUOC, a metric that jointly captures classification accuracy and ranking error while accounting for class imbalance and unobserved categories. It is computed by tracing paths along the confusion-matrix diagonal, with a Benefit term rewarding large correct entries and a Penalty term proportional to the distance from the diagonal. A perfect model will give an auoc of 0. AUOC is less penalizing when a fire event is predicted as a nearby but incorrect class (e.g., predicting class 0 instead of 1) and vice versa.

3. Intersection over Union (IoU), which measures how well the predicted risk aligns with actual risk when an event occurs. IoU is well-suited for multiclass wildfire prediction as it accounts for class uncertainty and preserves class ordinality—predicting class 1 instead of 4 is penalized less than predicting 0. With this metric, a false prediction is penalized proportionally to the class (0 to 4 is worst than 0 to 1 and vice versa).

4 Results

Tables 3 present the models performance for predicting FO and BA. Table 4 present the models performance for predicting fire risk in each department. This is give by equation 1 for T being each studied area (here French department), and score a score function.

$$Score = \frac{\int_0^T \text{score}(t)\,\mathrm{d}t}{\int_0^T \text{max\_score}\,\mathrm{d}t} \quad (1)$$

This score analyze how well the model is able to predict everywhere. The idea behind this equation is to compare the area under the curve representing each department's score, normalized by the maximum score (here 1 for each metric). By selecting the departments where at least one fire occurred in 2023, we can evaluate the model's ability to make predictions within each department. Note that this equation works when the score has a maximum value and this value is the best value possible. For that reason we don't compute the area score for auoc (minimum is better).

4.1 Model Performance on Global Metrics

Across all models and tasks, recall values are consistently higher than precision, both in global and area-normalized evaluations. For instance, in FO prediction, CatBoost reaches a recall of 0.51 versus a precision of 0.37, while GRU attains 0.56 recall for only 0.31 precision. This behavior is desirable, specially in low risk region where missing an actual fire is more critical than producing false alarms. We observe that FWI achieves the highest recall with the lowest AUOC, while the other metrics remain very low. This suggests that the metric, in its original implementation, has rather limited relevance for predicting occurrences

4.2 Model-Specific Observations

Among all tested models, CatBoost and GraphCastGRU achieve the best F1 scores (both 0.43). GraphCastGRU offers superior recall (0.60 vs. 0.33 precision), a higher IoU (0.25 compared to CatBoost's 0.24), and better spatial generalization. From the area-normalized evaluation, GraphCastGRU maintains F1 = 0.16, IoU = 0.09 for FO, and IoU = 0.08 for BA.

4.3 Comparison of FO and BA

When comparing the average IoU values of Fire Occurrence (FO) and Burned Area (BA), FO appears slightly easier to predict. Global IoU scores for FO range between 0.22–0.24 across models, compared to 0.21–0.23 for BA. This indicates that FO, as a target, is more predictable, whereas BA involves higher uncertainty linked to the dynamics of fire spread. There are several reasons that may explain this phenomenon: (1) Burned Area is inherently a more stochastic target compared to occurrence, as it depends not only on ignition but also on spread dynamics influenced by local conditions. (2) As observed earlier, the class distribution of Burned Area is more imbalanced than that of Fire Occurrence, leading models to underpredict higher BA classes and thus lowering performance metrics.

4.4 Binary vs. Multi-class Evaluation

Binary and multi-class models achieve relatively close F1 scores. For FO, binary CatBoost reaches F1 = 0.13 under area normalization, while multi-class CatBoost scores 0.13 as well. GRU binary and multi-class both yield F1 = 0.16. However, precision-recall trade-offs differ: binary classifiers tend to favor higher recall (e.g., 0.26 for GRU binary) at the cost of lower precision (0.13), while multi-class models remain more balanced (precision around 0.13–0.15, recall around 0.22–0.27).

This confirms that probability prediction alone is not sufficient at this scale: the ordinal structure of fire risk levels carries valuable information that can be better leveraged. Moreover, both global and area-normalized results show that further improvements require going beyond pure probability outputs.

Overall, these findings underline the necessity of multi-metric evaluation. Recall-oriented behavior is well aligned with operational needs, but only models such as GraphCast-GRU combine high F1 (0.43), strong IoU (0.25), positioning it as the most promising approach for localized fire risk prediction in France. The generalization accross department is a still a main challenge. Figure 3 shows predictions obtain by GraphCastGRU in Bouches du Rhône. Based on theses predictions we can clearly see the challenging task of predicting highest classes.

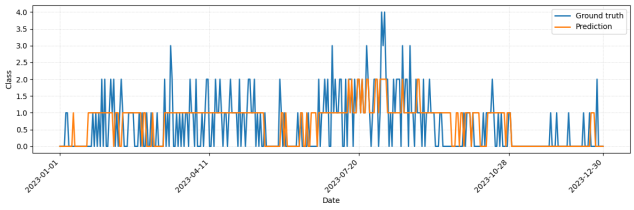


Figure 3: GraphCastGRU prediction for Bouches du Rhone in 2023.

Table 3: Model comparison for FO and BA at 0-day "LR" stands for Logistic Regression. We do not show FWI performance on BA as it is made for FO. In addition, IoU is not shown for binary models as it is not comparable with multi-classification. Best values are shown in bold.

Model	FO					BA	
	F1	Prec	Rec	IoU	auoc	IoU	auoc
FWI	0.17	0.10	**0.87**	0.06	**0.619**	-	-
LR	0.39 ± 0.01	0.28 ± 0.00	0.65 ± 0.01	0.23 ± 0.00	0.66 ± 0.00	0.24 ± 0.00	0.71 ± 0.00
catboost	**0.43 ± 0.00**	**0.37 ± 0.00**	0.51 ± 0.00	0.24 ± 0.00	0.72 ± 0.00	**0.25 ± 0.00**	0.71 ± 0.00
xgboost	0.42 ± 0.01	0.32 ± 0.01	0.60 ± 0.01	0.24 ± 0.00	0.70 ± 0.00	**0.25 ± 0.00**	0.71 ± 0.01
GRU	0.40 ± 0.01	0.31 ± 0.01	0.56 ± 0.04	0.23 ± 0.00	0.70 ± 0.01	0.23 ± 0.01	0.71 ± 0.01
LSTM	0.40 ± 0.01	0.29 ± 0.01	0.61 ± 0.01	0.23 ± 0.00	0.69 ± 0.01	0.23 ± 0.01	0.70 ± 0.01
MLP	0.39 ± 0.01	0.31 ± 0.02	0.52 ± 0.02	0.22 ± 0.01	0.71 ± 0.01	0.21 ± 0.01	**0.69 ± 0.00**
DilatedCNN	0.35 ± 0.01	0.24 ± 0.01	0.61 ± 0.02	0.20 ± 0.01	0.70 ± 0.00	0.21 ± 0.01	0.71 ± 0.01
graphCastGRU	**0.43 ± 0.01**	0.33 ± 0.02	0.60 ± 0.05	**0.25 ± 0.00**	0.70 ± 0.01	**0.25 ± 0.01**	0.71 ± 0.01
catboost Binary	**0.43 ± 0.00**	**0.37 ± 0.01**	0.50 ± 0.01	-	-	-	-
GRU Binary	0.40 ± 0.01	0.30 ± 0.01	0.62 ± 0.01	-	-	-	-
NetMLP Binary	0.39 ± 0.01	0.30 ± 0.01	0.55 ± 0.02	-	-	-	-

Table 4: Area model comparison for FO and BA at 0-day prediction.

Model	FO				BA
	F1	Prec	Rec	IoU	IoU
FWI	0.14	0.08	**0.83**	0.058	-
LR	0.15 ± 0.01	0.14 ± 0.01	0.29 ± 0.01	**0.09 ± 0.00**	0.08 ± 0.00
catboost	0.13 ± 0.00	0.13 ± 0.00	0.18 ± 0.00	0.07 ± 0.00	0.08 ± 0.00
xgboost	0.15 ± 0.00	0.13 ± 0.01	0.24 ± 0.00	**0.09 ± 0.00**	0.08 ± 0.01
GRU	**0.16 ± 0.02**	0.13 ± 0.02	0.24 ± 0.04	**0.09 ± 0.00**	0.08 ± 0.00
LSTM	**0.16 ± 0.01**	0.13 ± 0.01	0.27 ± 0.01	**0.09 ± 0.00**	**0.09 ± 0.01**
MLP	0.15 ± 0.01	**0.15 ± 0.01**	0.22 ± 0.02	0.08 ± 0.01	**0.09 ± 0.00**
DilatedCNN	**0.16 ± 0.01**	0.12 ± 0.01	0.31 ± 0.03	**0.09 ± 0.01**	**0.09 ± 0.01**
graphCastGRU	**0.16 ± 0.00**	0.14 ± 0.01	0.23 ± 0.04	**0.09 ± 0.00**	0.08 ± 0.01
catboost Binary	0.13 ± 0.01	0.13 ± 0.00	0.18 ± 0.00	-	-
GRU Binary	**0.16 ± 0.00**	0.13 ± 0.01	0.26 ± 0.01	-	-
NetMLP Binary	**0.16 ± 0.01**	**0.15 ± 0.02**	0.24 ± 0.01	-	-

5 Features importance

We investigated feature importance by computing SHAP values on the multi-class CatBoost model. Figure 4 shows the top 15 features for FO. A corresponding figure for BA prediction is provided in the supplementary materials.

Fire prediction is primarily driven by historical and short-term signals. Encoded temporal variables, spatial clustering, and immediate weather conditions dominate the model's decision-making. Fire danger indices, relative humidity, and vegetation-specific factors—such as the presence of pine trees—also contribute meaningfully.

Historical features are strong predictors in historically active regions but can hurt generalization in low-risk or emerging areas. While they help the models to easily detect the seasonal trend of fire, their static nature biases the model toward past patterns, potentially missing new fire-prone zones.

6 Discussion

In this section, we outline several possible directions for future work aimed at improving current model performance.

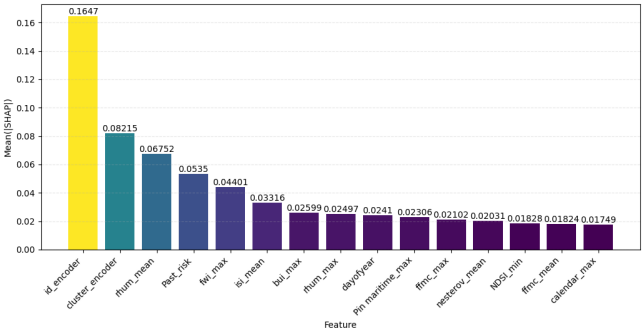


Figure 4: Top 15 features computed on multi-classification Catboost model on different time horizons for FO. ID encoder corresponds to Department ID

- Clustering Current risk classes are based solely on historical FOs and BA at the departmental level. Incorporating seasonal data and clustering similar regions may help capture risk dynamics that go beyond purely historical patterns.

- Ordinal Classification Training Risk levels are inherently ordered, yet current models treat them as independent categories. Future work could leverage ordinal-aware loss functions (e.g., Kappa loss) to exploit this structure and better handle class imbalance.

- Federated Learning A single global model may overlook important regional patterns. Federated learning would allow each region or cluster to train a specialized model while still benefiting from shared knowledge to enhance generalization. This approach would also facilitate the inclusion of overseas departments and Corsica, which have been excluded thus far due to their significantly different vegetation and meteorological conditions compared to mainland France.

- Filtering Features Our analysis of feature importance revealed that historical variables are strong predictors in regions with frequent past fire activity. While these features help the model capture seasonal fire trends, their static nature may bias predictions toward past patterns and overlook emerging fire-prone areas. Reducing their influence during training could help the model prioritize more transferable features. We also acknowledge that 2D convolutional networks such as ResNet and ConvLSTM were not included in our benchmark. Although effective for spatial tasks like fine-scale map prediction, their high computational cost made them unsuitable for our department-scale feature set. We plan to explore 2D CNNs (after dimensionality reduction) in future work.

7 Conclusion

In this work, we introduced a new nationwide database for wildfire risk prediction, specifically aligned with the departmental organization of French firefighting services and enriched with local variables such as calendar features. While this dataset can serve broader research domains, we proposed the first benchmark for predicting both fire occurrence and burned area using an ordinal classification scheme, a novel approach in this context. Our experiments highlight three key findings: (1) multi-class prediction provides a clear advantage over binary formulations at this scale, (2) GraphCast-GRU achieves the best overall performance across metrics which corroborate with findings of [6], and (3) the current ordinal scheme is less suitable for burned area prediction. Future directions include exploring ordinal-aware loss functions, applying federated learning to assess generalization, and testing more computationally demanding architectures such as CNNs and Transformers to better capture spatial and temporal dependencies.

References

[1] Suzanne E Finlay, Andrew Moffat, Rob Gazzard, David Baker, and Virginia Murray. Health impacts of wildfires. *PLoS Currents*, 4:e4f959951cce2c, Nov 2012. .

[2] Hongyu Chen, Jonathan M Samet, Philip A Bromberg, and Haiyan Tong. Cardiovascular health impacts of wildfire smoke exposure. *Particle and Fibre Toxicology*, 18(1):2, Jan 2021. .

[3] Wayne E. Cascio. Wildland fire smoke and human health. *Science of the Total Environment*, 624:586–595, May 2018. ISSN 1879-1026. .

[4] Spyros Kondylatos, Ioannis Prapas, Michele Ronco, Ioannis Papoutsis, Gustau Camps-Valls, Maria Piles, Miguel-Angel Fernandez-Torres, and Nuno Carvalhais. Wildfire danger prediction and understanding with deep learning. *Geophysical Research Letters*, 49 (17):e2022GL099368, 2022. . URL . e2022GL099368 2022GL099368.

[5] Ilektra Karasante, Lazaro Alonso, Ioannis Prapas, Akanksha Ahuja, Nuno Carvalhais, and Ioannis Papoutsis. Seasfire as a multivariate earth system datacube for wildfire dynamics, 2023. URL .

[6] Dimitrios Michail, Charalampos Davalas, Lefki-Ioanna Panagiotou, Ioannis Prapas, Spyros Kondylatos, Nikolaos Ioannis Bountos, and Ioannis Papoutsis. Firecastnet: Earth-as-a-graph for seasonal fire prediction, 2025.

[7] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Graphcast: Learning skillful medium-range global weather forecasting, 2023.

[8] Dimitris Sykas, Dimitris Zografakis, Konstantinos Demestichas, Constantina Costopoulou, and Pavlos Kosmidis. Eo4wildfires: An earth observation multi-sensor, time-series machine-learning-ready benchmark dataset for wildfire impact prediction, March 2023. URL .

[9] Spyros Kondylatos, Ioannis Prapas, Gustau Camps-Valls, and Ioannis Papoutsis. Mesogeos: A multi-purpose dataset for data-driven wildfire modeling in the mediterranean, 2023. URL .

[10] Lara Vilar, Douglas G. Woolford, David L. Martell, and M. Pilar Martín. A model for predicting human-caused wildfire occurrence in the region of madrid, spain. *International Journal of Wildland Fire*, 19(3): 325–337, 2010. . URL .

[11] François Pimont, Héléne Fargeon, Thomas Opitz, Julien Ruffault, Renaud Barbero, Nicolas Martin-StPaul, Eric Rigolot, Miguel RiviÉre, and Jean-Luc Dupuy. Prediction of regional wildfire activity in the probabilistic bayesian framework of firelihood. *Ecological Applications*, 31(5):e02316, 2021. .

[12] J. Chen, X. Wang, Y. Yu, X. Yuan, X. Quan, and H. Huang. Improved prediction of forest fire risk in central and northern china by a time-decaying precipitation model. 2022. URL .

[13] M Meinard Muller. *Dynamic Time Warping*, pages 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-74048-3.

[14] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[15] Sylvain Dupire, Thomas Curt, and S. Bigot. Spatio-temporal trends in fire weather in the french alps. *Science of The Total Environment*, 595:801–817, 10 2017. .

[16] Wilson Silva, João Ribeiro Pinto, and Jaime S. Cardoso. A uniform performance index for ordinal classification with imbalanced classes. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018. .