A Voting System to Optimize Daily Forest Fire Prediction

1st Nicolas Caron
FEMTO-ST Institute
Université Marie et Louis Pasteur
Besançon, France
nicolas.caron@univ-fcomte.fr

2nd Hassan Noura
FEMTO-ST Institute
Université Marie et Louis Pasteur
Besançon, France
hassan.noura@univ-fcomte.fr

3rd Christophe Guyeux
FEMTO-ST Institute
Université Marie et Louis Pasteur
Besançon, France
christophe.guyeux@univ-fcomte.fr

4th Benjamin Aynes
SAD Marketing Company
Lille, France
b.aynes@sad-marketing.com

***Abstract*—Forest-fire prediction using Artificial Intelligence (AI) continues to face major challenges, including (i) the ability to generalize across regions with very different risk profiles, (ii) managing the inherent daily variability and randomness of fire occurrences (including extreme fire days). These factors together have hindered the deployment of dependable prediction systems in operational settings. In this work, we introduce a novel multi-risk modeling framework specifically designed to tackle all two challenges simultaneously. The proposed approach is applied to daily forest-fire prediction across mainland France. We develop a voting-based system that combines the outputs of multiple models trained on signals smoothed with a range of convolutional kernels, capturing both local and seasonal variations. The proposed solution achieves superior performance compared to conventional models, demonstrating improved cross-regional transferability and robustness to daily fluctuations. Notably, it significantly enhances prediction skill for the rare but damaging extreme-fire days, where traditional models often fail. Our experiments reveal that using an ensemble of multiple risk models can better capture the complex dynamics of fire risk and provide more reliable guidance for decision-makers. Supplementary materials are available here.**

***Index Terms*—Machine learning, Wildfire prediction, Extreme prediction, Voting**

I. Introduction

Forest fires impose extensive costs on multiple fronts. Economically, the immediate destruction they wreak on property and infrastructure is evident. According to the U.S. National Interagency Fire Center, the U.S. alone spent over $3 billion on wildfire suppression in 2018 without account post-fire costs related to rebuilding, loss in property values, and disruptions to local economies [1]–[3]. According to AccuWeather, the economic loss from the large Los Angeles fire of 2025 is estimated at $250 billion. From a human perspective, in addition to the immediate threat to life, wildfires cause long-term health impacts due to the inhalation of smoke and fine particulate matter. For instance, a study found that the 2018 Camp Fire in California resulted in over $16.5 billion in health costs from air pollution exposure alone [4]. Ecologically, fires can decimate local flora and fauna, with some species taking decades, if not centuries, to recover. The loss of vegetation further exacerbates the risk of soil erosion and disrupts the local water cycle, posing long-term ecological risks [5]. Moreover, the carbon released during wildfires contributes significantly to global greenhouse gas emissions, thus amplifying the effects of climate change.

A. Problem formulation

According to Jan-Miguel-Ayanz et al [6], wildfire risk can be conceptualized as the likelihood of a fire event's occurrence and the potential adverse consequences arising from it. Developing an algorithm to predict this likelihood offers an essential tool in proactive wildfire prevention. The objective of this article is to provide fine-grained, operational forecasts of the daily number of wildfires at short horizons (daily) in French departments. This enables better risk management, intervention planning, and protection of populations and infrastructure.

B. State of the art

Table I shows the analysis of the state of the art carried out in this article. Our literature review highlighted several important points: daily predictions, although suggested in some papers, are not achieved and suffer from randomness issues, especially in areas with high occurrence rates. A single model struggles to handle multiple regions, each with different risk levels.

C. Contributions

In this article, we propose a method to optimize the daily prediction of the number of fires for a given region. We address two challenges highlighted in the state of the art: handling regions with different risk levels and the random nature of fires. Using a public database, we demonstrate that a single risk model is not the ideal solution. Our method involves applying a series of convolutional filters to the raw signal and implementing a voting system. Our results show that, depending on the specific task (predicting the highest risk, predicting in a particular area, predicting during a specific season, etc.), the optimal number of models varies. This analysis suggests

TABLE I
SUMMARY OF RECENT WILDFIRE–OCCURRENCE PREDICTION STUDIES

Authors / Citation	Objective	Remarks
[7]–[9]	"Daily" wildfire-danger modelling from daily covariates.	Archive converted to isolated fire / non-fire pixels; non-fire points chosen by distance rule (e.g. >10 km). Random 80/20 split yields high apparent skill but ignores day-to-day uncertainty, inducing a saw-tooth signal.
Bethany et al. [10]	Predict daily fire-occurrence risk maps.	Accuracy rises as evaluation windows widen from 40 km to 120 km—coarser grids absorb spatial noise. Global performance low due to stochastic nature.
Graff et al. [11]	Daily ignition probability via Poisson regression.	New-fire occurrence is highly random at a one-day horizon; MLPs confer no gain over the simpler model.
[12]–[14]	Monthly fire-probability forecasting.	Enlarging the temporal scale smooths the stochastic nature of ignition events.
Michail et al. [15]	Weekly global fire-occurrence prediction with GraphCast + temporal encoder.	One week ahead, AUPRC = 0.64 worldwide, but only 0.20 for Europe; a single "global" network hides pronounced regional disparities, motivating domain-specific fine-tuning.

the need for a complex framework composed of multiple risk prediction models. Figure 1 shows the pipeline developed in this article.

D. Organization

First, we describe the database employed, detailing the predictor variables and the target response to be estimated. Next, we examine the proposed voting scheme in depth. We then present the training configuration for each of the tested models and report the corresponding experimental results. Finally, we offer guidelines for kernel selection aimed at optimizing a specific downstream task.

II. Dataset

This section presents and discusses the data source used to obtain fire information. It describes the features and the target variable to be predicted.

A. The Forest Fire Database

Also called BDIFF, it is an online 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. This database provides access to multiple sources of information about forest fires, as follows:

- Date of the day
- Time the alert was issued
- The nearest city
- The cause (criminal, accidental, etc.), when available
- The burned area.
- Meteorological data

We used data from 2017 to 2023.

B. Features and Target

Features and target were computed as it is described in the unpublished article of Caron et al. [16] (see figure 2).

The wildfire prediction problem is treated as an ordinal multi-class classification task. For each department, an ordinal 5-class signal for occurrence using the K-Means algorithm is created. Each sample with 0 fires was set to class 0, and positive samples were clustered into four classes, representing Normal, Medium, High, and Extreme fire occurrence. This approach highlights typical occurrences rather than absolute values.

The features computed for training the models are grouped into 4 categories, as shown in Table II: Meteorological, Topographic, Socio-Economic, and Historical. The feature processing follows the same procedure as noted in the original article: data are transformed into a 3D raster with a 2 km resolution, aggregated by department using the mean, maximum, and minimum, and features with low variance or high correlation are removed using Pearson, Spearman, and Kendall coefficients (keeping the highest variance).

The dataset is split into training (2017–2021), validation (2022), and testing (2023) subsets. All calculations were performed on the training set and generalized to all data. Each row consists of the aggregated features and the target risk for a specific cluster within a specific French department.

III. Proposed Voting Scheme

In this section, we present the proposed voting scheme to optimize forest fire prediction. This idea stems from the desire to address the randomness inherent in fire signals. A fire is a rare event that can occur due to multiple factors that are difficult to account for (such as cigarette butts, barbecues, electrical cables...), not to mention fires of criminal origin. Therefore, a fire occurring on day J could reasonably have happened within a certain range of days around J (from J-x to J+x), where x is an integer. This phenomenon creates discrete datasets in which some samples labeled 0 do not correspond to a day without risk. When focusing not only on the occurrence of fires but also on the number of fires, the signal becomes increasingly noisy. This effect is even stronger when the study region (here, France) is large. In some departments (e.g., Ain), very few forest fires have occurred (11 over 7 years), yet vegetation dynamics could make these departments much more at risk than current data suggests. However, predicting events with only 11 samples is challenging. Among these rare events, extreme events (many fires in a single day) are even harder to predict.

Figure 3 shows the final solution proposed to bypass those limitations. Starting from the fire signal, different convolution

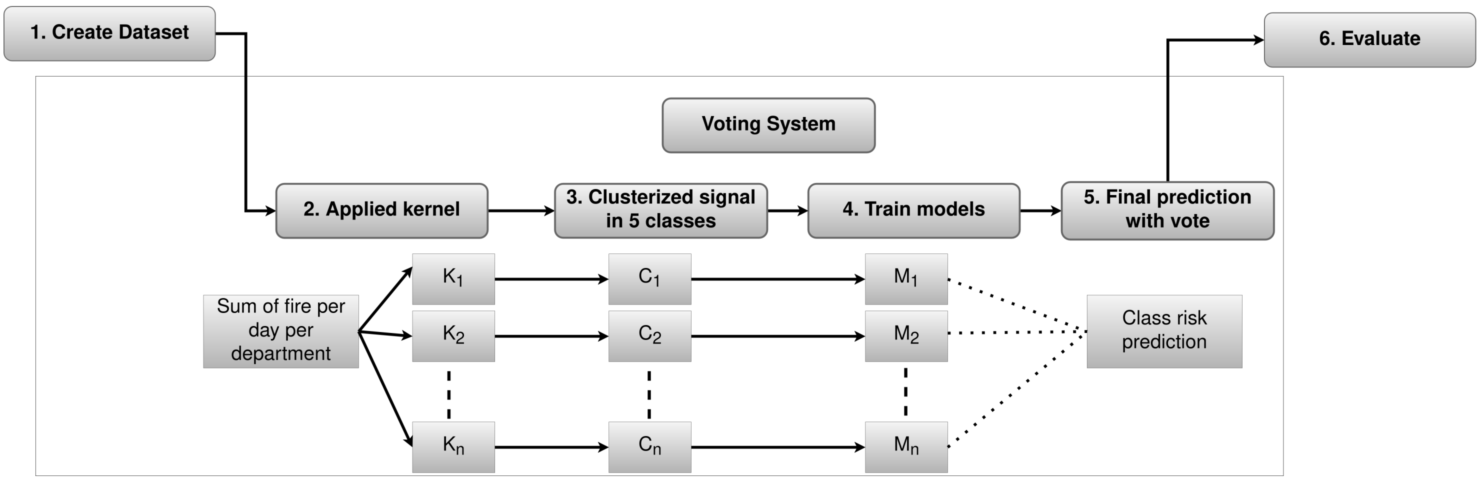

Fig. 1. Process pipeline. After creating the dataset, different kernels are applied through convolution or simple aggregation. This resulted in different signals, for each of which we trained a model. The final prediction is given by the votes of all models.

TABLE II
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	IGN
Dew Point	-	-	Forest landcover	-	-
Precipitation	-	-	Landcover	-	Cosia
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	OSMNX [17]	Past risk	Daily	Calculated
Population	-	Kontur	Past risk burned area	-	-
Calendar	Daily		Cluster	Static	-
			Department	-	-

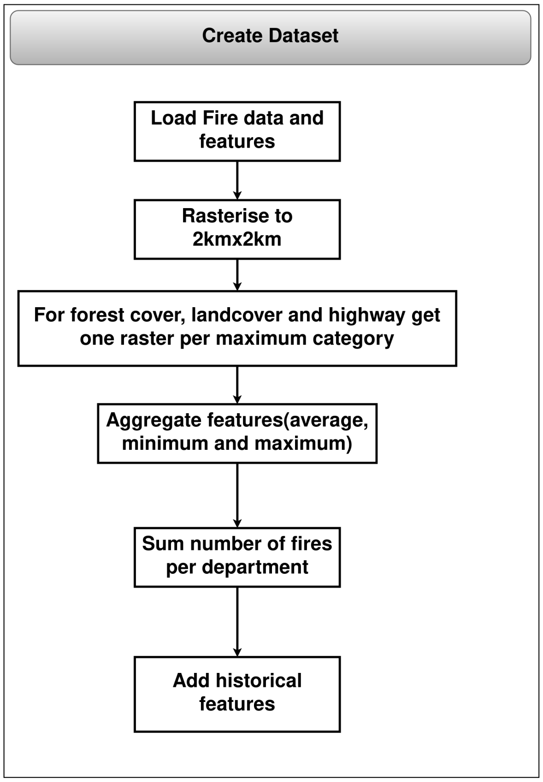

Fig. 2. Dataset process employed in this article.

kernels (mean, max, min, cubic, quartic, etc.) are applied to smooth the original signal. Each filtered output is clustered using the same method as the raw signal and then passed to an AI model. In the end, one model is obtained for each applied kernel. The final output is then the result of a vote among all models.

The kernels used, as well as their sizes (in number of days), can be difficult to define without prior training. For that reason, we selected a large number of kernels in order to maximize the information extracted. Figure 4 shows all the kernels studied. Kernel size range is defined in

$$size \in \{1, 3, 5, Specialized\},$$

Numerical data defines the number of days before and after. The specialized size is computed per season and cluster. The size corresponds to the mean duration of continuous fire sequences (consecutive fire days with gaps of less than 2 days between them). A 3-day inactivity threshold separates sequences. For each signal (including the original), we created 5 classes (representing: Null, Low, Medium, High, and Extreme risk) using K-Means.

This process results in a voting scheme with 32 different models plus a model trained on the raw clustered fire signal for a total of 33 models.

There are several ways to handle voting among AI models. In this article, we use a soft voting approach based on each model's class-membership probabilities. The vote is weighted by the model's performance on the validation set, calculated on the original (unfiltered) target fire signal.

IV. Training

To evaluate our solution, several models have been used. The tested models are listed in Table III with their configurations provided in the supplementary materials. In addition to the classic versions, CatBoost, logistic regression, and XGBoost were also trained with oversampling using SMOTE. Three oversampling rates were tested: 2×, 4×, and 6×. Oversampling was applied only to the original fire signal (with no kernels). For each training, we tested different percentages of 0 samples (ranging from 0.05 to 1.0), selecting the one that gave the best score on the validation set.

The high training time and the size of the framework prevent us from using more complex and heavy models, such as GNNs or CNNs. However, it is worth noting in the unpublished article [16] that very complex models did not significantly improve performance compared to simple models (for this dataset). While this is planned for future work, our objective here is to demonstrate that a single risk system cannot address the aforementioned limitations.

TABLE III
List of Models Used.

Model	**Type**
LG (Logistic Regression)	Linear model
XGBoost	Decision tree (Boosting)
CatBoost	Decision tree (Boosting)
GRU	Recurrent neural network
LSTM	Recurrent neural network
MLP	Fully connected neural network

V. Results

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

1) Binary f1 score.
2) 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. Unlike metrics such as Balanced Class Accuracy, IoU reflects the severity of misclassifications. It also highlights model performance on rare events, where accurate prediction often requires accepting more false positives—a trade-off that IoU captures effectively.

In this paper, we only show the performance on the IoU metric, as it is more precise than the F1 score, which is provided in the supplementary materials.

Each metric was calculated using three methods:

- **Classic**: computed over the entire dataset.
- **Generalization**: evaluates the area under the score curve for each department as it was used in [16].

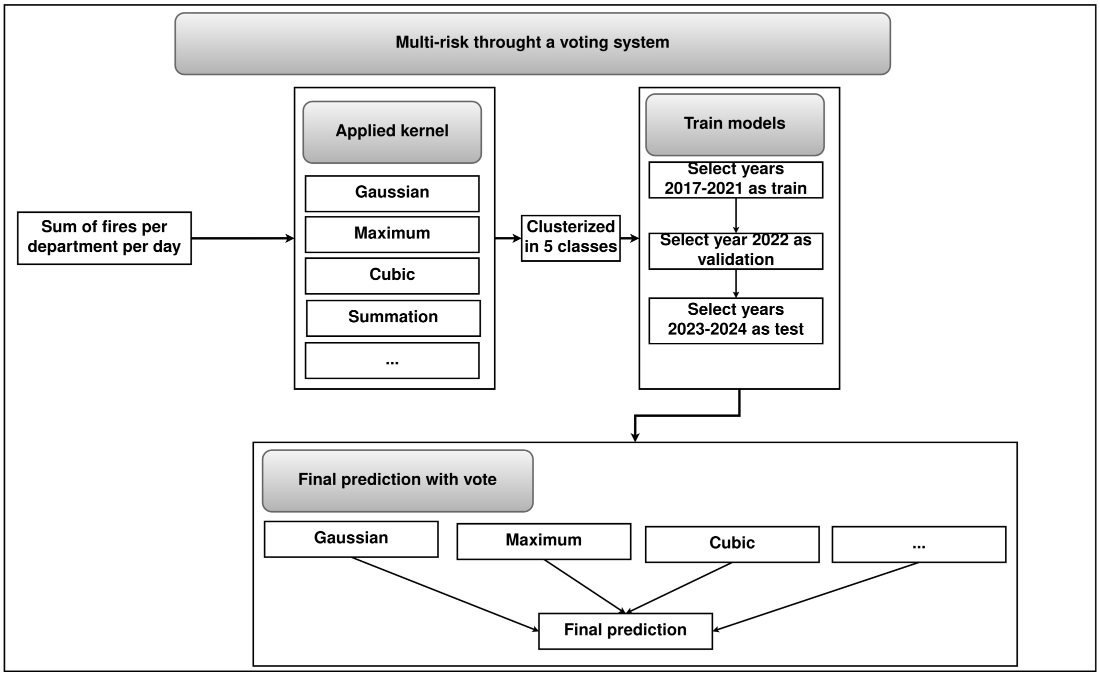

Fig. 3. The voting system built in this article to optimize wildfire daily prediction.

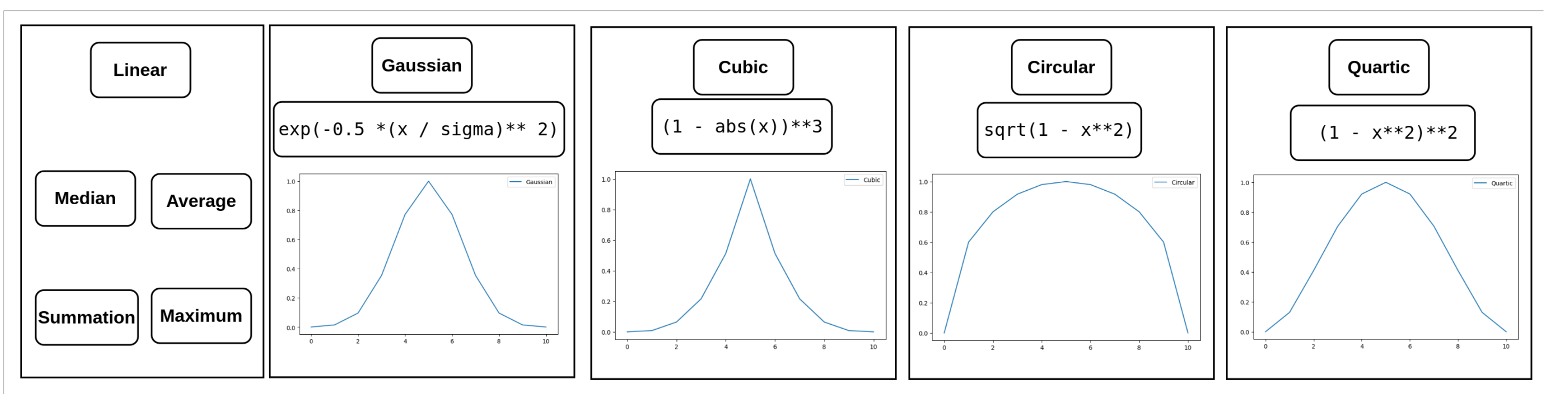

Fig. 4. Types of kernels applied to the fire signal. Sigma was set to kernel_size - 1.

- **Extreme**: evaluates the Intersection over Union (IoU) on the higher classes (2, 3, and 4).

Using the validation-set scores (used as voting weights), we tested the voting system while varying the number of models included—from 1 up to 20, and then with the entire set of models.

A. Classic evaluation

The overall performance of the models is presented in figure 5. As can be seen, in the case of XGBoost and LG, the voting models slightly improved the results ($\sim +0.003$). In deep learning models (MLP, GRU, and LSTM), we notice a larger improvement in the global scores using the voting schemes, but both performances remain similar. For CatBoost, no noticeable improvement is observed. The use of SMOTE reduced the models' performance (~ -0.03).

B. Generalization evaluation

This section investigates the models' generalization scores—their ability to cope with different regions (here, individual French departments) that have distinct risk histories. Using Equation 1, the score is first computed for each department (t), then normalized by the maximum possible area (which equals 1 for each department). Only departments that experienced at least one fire during the 2023 test year are considered.

Figure 6 displays the resulting scores. Although applying SMOTE with an oversampling factor of 6 improved the score,

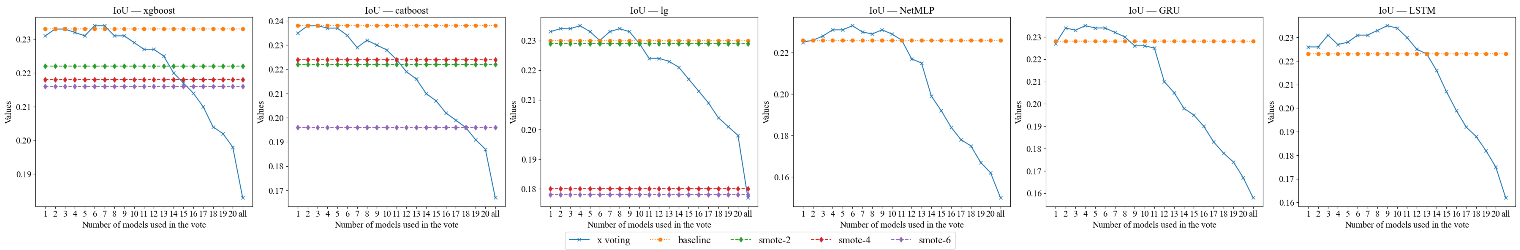

Fig. 5. IoU performance on full database between voting, classic, and SMOTE models.

the dedicated scoring strategy achieved the best performance (0.094 with CatBoost using an ensemble of ten models). Nevertheless, the score remains very low, highlighting the persistent difficulty of the task. A similar pattern is observed across the other models evaluated. However, although the GRU voting model achieves its maximum score with the same number of models (around 10), it did not surpass the generalization of the baseline model, even if the scores are very close (0.094 vs 0.095).

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

C. Extreme events evaluation

This section evaluates model performance on the most extreme days. To do so, two separate sample groups were analysed: (i) all observations belonging to classes 2, 3, and 4, and (ii) all observations belonging to classes 3 and 4.

Figure 7 reports the results for the first group. When the number of models in the voting ensemble is greatly increased to 12-14, the voting scheme achieve higher precision than with the baseline models and those trained with SMOTE. When comparing performance with generalization enabled, we observe a correlation between the improvement in predictions for classes superior than 2 and the generalization score: the former becomes noticeable from six models onward, and the latter from seven. This suggests that generalization particularly enhances the prediction of abnormal days in low-risk departments.

Figure 8 shows the results for the second group. Here the improvement is even stronger: for CatBoost, the score rises from 0.341 to 0.402 when the full ensemble is used. Logistic regression follows a similar trend but drops sharply when all models vote, which suggests greater stability at around 20 models. The greatest improvement in extreme values is achieved by the GRU model, increasing from 0.292 to 0.415.

Applying SMOTE does not yield strong stability across the models. By contrast, the kernel-based method produces more reliable high-class values because it embeds seasonal information via the different kernel types. Instead of creating synthetic points, it identifies new days that are potentially at risk.

VI. Selecting Best Kernels for a Specific Task

Validation scores obtained by each model, which are used as weights in the votes, are shown in Figure 9. This figure illustrates a proposed segmentation based on the results analyzed previously. From this segmentation, we identify specific filters tailored to each task:

- **Median filter** ((highest weight): produces classifications that closely mirror the raw signal, regardless of the filter size.
- **Linear aggregations** (max, sum): improve spatial generalization by smoothing out the randomness of fires in high-risk regions and emphasizing seasonal risk patterns rather than daily fluctuations.
- **Complex kernels** (Gaussian, cubic, circular, quartic), when scaled according to season and region, yield substantial gains in extreme-risk prediction while maintaining primary emphasis on the day of filter application (the kernel center).
- **Additional filters**: In light of the results, the following models were not further analyzed, and we found little benefit in including them going forward.

VII. Conclusion

In this article, we proposed a method to address the generalization issue in low-risk regions and the stochastic nature of daily fire predictions. Our approach creates several *ordinal risk systems* and assigns to each day the risk level that receives the most votes across the trained models. We found that, depending on the task, the number of models considered can vary: selecting only the best-performing models tends to improve overall predictions, while including a larger ensemble improves the prediction of extreme values and enhances generalization. Until now, risk systems have been built by applying different types of filters to the fire signal. The proposed method delivers better performance and flexibility compared to conventional over-sampling. The main drawbacks are (i) the training time required for the full ensemble (which may take several days, depending on the model and hardware), and (ii) the final model size. For the latter, we believe a

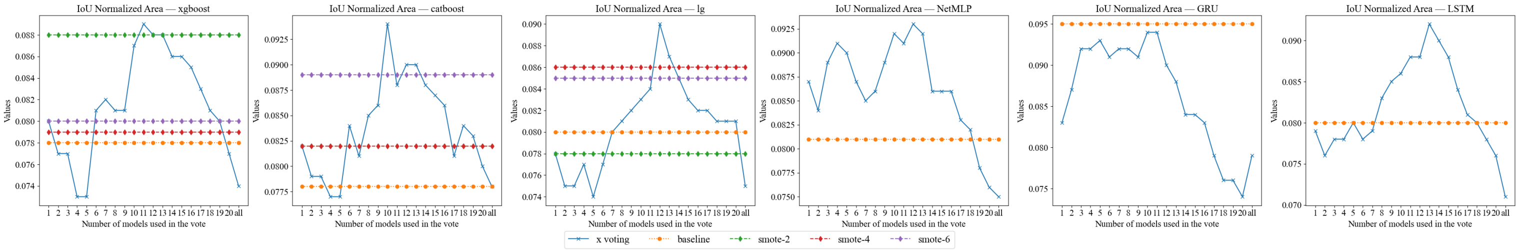

Fig. 6. Area IoU performance on full database between voting, classic, and SMOTE models.

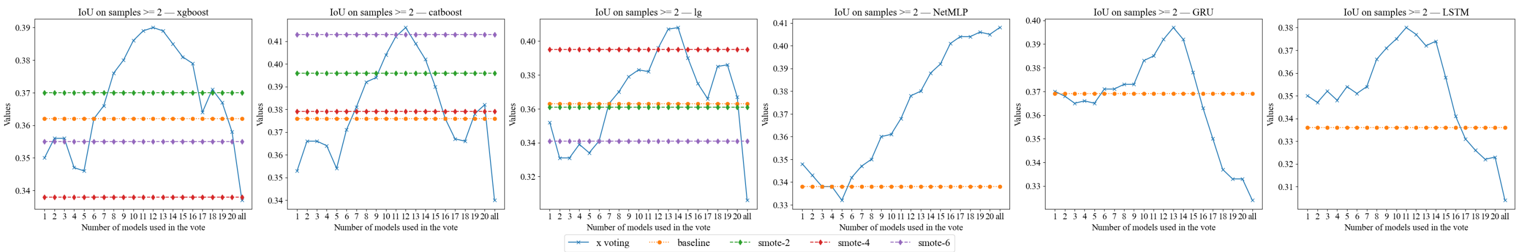

Fig. 7. IoU performance on samples of true risk (or predicted) on samples of class superior or equal to 2 between voting, classic, and SMOTE models.

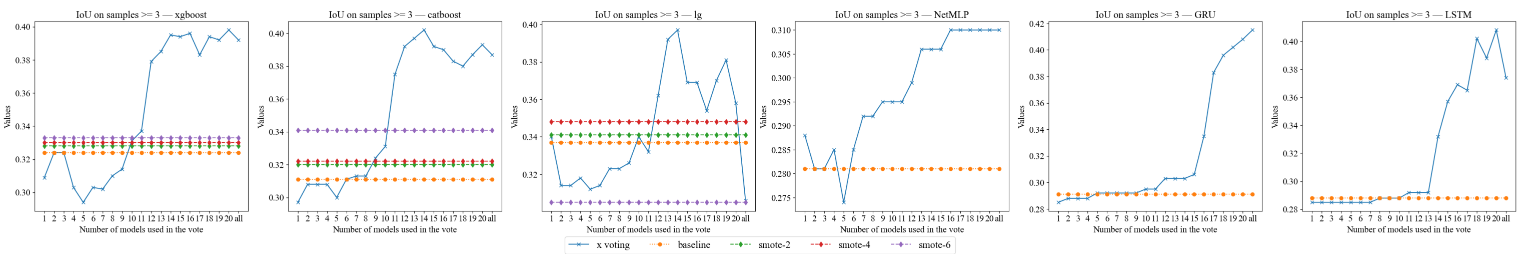

Fig. 8. IoU performance on samples of true risk (or predicted) on samples of class superior or equal to 3 between voting, classic, and SMOTE models.

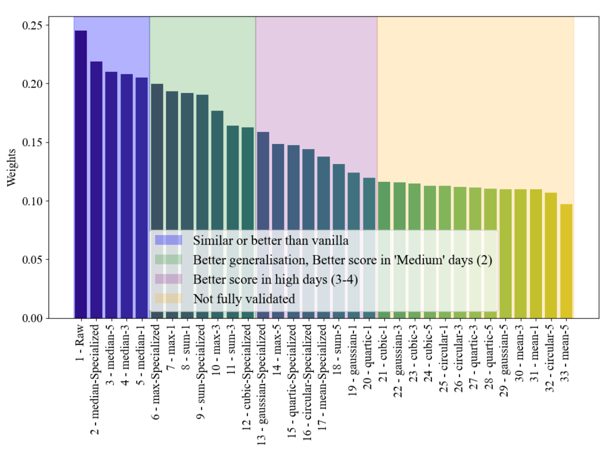

Fig. 9. Average score obtained by each model trained on a derived signal. X axis is formatted as position-kernel-size.

distillation-based strategy could shrink the footprint: the task-specific ensemble would act as the *teacher* and a smaller network would serve as the *student*.

Acknowledgment

All computations have been performed on the Mésocentre de calculs de Franche-Comté supercomputer facilities.

Declaration

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Wang, D. Guan, S. Zhu, and et al., "Economic footprint of california wildfires in 2018," *Nature Sustainability*, vol. 4, no. 3, pp. 252–260, 2021. [Online]. Available: https://doi.org/10.1038/s41893-020-00646-7

[2] J. M. Diaz, "Economic impacts of wildfire," *Southern Fire Exchange*, vol. 498, pp. 2012–7, 2012.

[3] M. W. Jones, J. T. Abatzoglou, S. Veraverbeke, N. Andela, G. Lasslop, M. Forkel, A. J. P. Smith, C. Burton, R. A. Betts, G. R. van der Werf, S. Sitch, J. G. Canadell, C. Santín, C. Kolden, S. H. Doerr, and C. Le Quéré, "Global and regional trends and drivers of fire under climate change," *Reviews of Geophysics*, vol. 60, no. 3, p. e2020RG000726, 2022. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020RG000726

[4] N. Fann, B. Alman, R. Broome, G. Morgan, F. Johnston, G. Pouliot, and A. Rappold, "The health impacts and economic value of wildland fire episodes in the u.s.: 2008–2012," *The Science of the total environment*, vol. 653, pp. 427–438, 2019.

[5] G. Certini, “Effects of fire on properties of forest soils: a review,” *Oecologia*, vol. 143, no. 1, pp. 1–10, 2005.
[6] J. San-Miguel-Ayanz, C. D, M. Alexander, K. Tolhurst, G. Morgan, R. Sneeuwjagt, and M. Dudley, *Current Methods to Assess Fire Danger Potential*, 01 2004.
[7] S. Kondylatos, I. Prapas, M. Ronco, I. Papoutsis, G. Camps-Valls, M. Piles, M.-A. Fernandez-Torres, and N. Carvalhais, “Wildfire danger prediction and understanding with deep learning,” *Geophysical Research Letters*, vol. 49, no. 17, p. e2022GL099368, 2022, e2022GL099368 2022GL099368. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022GL099368
[8] C. Wang, H. Liu, Y. Xu, and F. Zhang, “A forest fire prediction framework based on multiple machine learning models,” *Forests*, vol. 16, no. 2, 2025. [Online]. Available: https://www.mdpi.com/1999-4907/16/2/329
[9] Z. Zhang, Y. Tian, G. Wang, C. Zheng, and F. Zhao, “A forest fire prediction method for lightning stroke based on remote sensing data,” *Forests*, vol. 15, no. 4, 2024. [Online]. Available: https://www.mdpi.com/1999-4907/15/4/647
[10] B. L. Earnest, A. McGovern, C. Karstens, and I. Jirak, “Part i: Improving wildfire occurrence prediction for conus using deep learning and fire weather variables,” *Artificial Intelligence for the Earth Systems*, vol. 3, no. 3, p. e230057, 2024. [Online]. Available: https://journals.ametsoc.org/view/journals/aies/3/3/AIES-D-23-0057.1.xml
[11] C. A. Graff, S. R. Coffield, Y. Chen, E. Foufoula-Georgiou, J. T. Randerson, and P. Smyth, “Forecasting daily wildfire activity using poisson regression,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4837–4851, 2020.
[12] S. Choi, M. Son, C. Kim, and B. Kim, “A forest fire prediction model based on meteorological factors and the multi-model ensemble method,” *Forests*, vol. 15, no. 11, 2024. [Online]. Available: https://www.mdpi.com/1999-4907/15/11/1981
[13] J. Koh, “Gradient boosting with extreme-value theory for wildfire prediction,” *Extremes (Boston)*, vol. 26, no. 2, pp. 273–299, Jan 21 2023, epub ahead of print.
[14] D. Cisneros, J. Richards, A. Dahal, L. Lombardo, and R. Huser, “Deep graphical regression for jointly moderate and extreme australian wildfires,” *Spatial Statistics*, vol. 59, p. 100811, 2024.
[15] D. Michail, C. Davalas, L.-I. Panagiotou, I. Prapas, S. Kondylatos, N. I. Bountos, and I. Papoutsis, “Firecastnet: Earth-as-a-graph for seasonal fire prediction,” 2025.
[16] N. Caron, C. Guyeux, H. Noura, and B. Aynes, “Localized forest fire risk prediction: A department-aware approach for operational decision support,” 2025. [Online]. Available: https://arxiv.org/abs/2506.04254
[17] G. Boeing, “Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks,” *Computers, Environment and Urban Systems*, vol. 65, pp. 126–139, 2017.