
Scalable & explainable ML for wildfire risk modeling in Southern Europe: A case-study in Portugal

Ophélie Meuriot

Denmark Technical University
ophme@dtu.dk

Jorge Soto Martin

Denmark Technical University
jsoma@dtu.dk

Beichen Zhang

Lawrence Berkeley National Laboratory
bzhang5@lbl.gov

Francisco Camara Pereira

Denmark Technical University
camara@dtu.dk

Martin Drews

Denmark Technical University
mard@dtu.dk

Abstract

Wildfires constitute one of the most devastating natural disasters, with wide impacts across economic sectors and society in Europe. The study shows how data-driven models can be used to combine human, topographic, land cover and climate data to quantify wildfire risk. Four machine learning models are trained on a historical record of fires (from June to October between 2008 and 2023) in Southern Europe excluding Portugal. The Random Forest (RF) model ($AUC = 0.91$ and $F1 = 0.85$) is then run daily in Portugal (12 x 12 km grid) during October 2017, a month during which devastating wildfires were the cause of 51 casualties. The model is found to represent high-risk areas more accurately than the widely used Fire Weather Index (FWI). Key features influencing the high fire risk during the period are identified using explainable AI. This study provides a scalable and lightweight model, which can be used to support climate impact assessments by (1) quantifying high-risk areas and (2) identifying key drivers to inform adaptation strategies.

1 Introduction

In the recent European Climate Risk Assessment (European Environment Agency, 2024), wildfires are identified as one of the most critical climate risks in Europe, with confirmed impacts on critical infrastructure, health, and other key sectors. Modeling wildfire risk remains a complex challenge, particularly as predictions of fire risk often rely on operational weather indices (Shaples, 2022) which exclude key factors such as human activities and other ignition sources. This limitation has led to ongoing debates about their effectiveness in predicting fire activity (Castel-Clavera et al., 2025). With the growing availability of remotely sensed data, the use of machine learning (ML) to assess wildfire risk has expanded, enabling the integration of diverse data sources (Bountzouklis et al., 2023; Oliveira et al., 2012; Dorph et al., 2022). For example, Oliveira et al. (2012) use a combination of climatic, land cover, and human activity to quantify wildfire risk in the Mediterranean.

So far, most data-driven studies have focused on time-invariant susceptibility maps (Shafapourtehrany, 2023; Erni et al., 2024; Iban and Aksu, 2024). With the growing need to anticipate and respond to the threat of wildfires, it is crucial to quantify the risk of wildfire under specific weather conditions. In addition, understanding the key local drivers of wildfire risk are critical to develop successful

adaptation measures. The aim of the study is to (1) develop a data-driven wildfire risk model for Southern Europe which combines factors, such as climate variables, land use, human activity and topography, to estimate the probability of wildfire occurrence for given daily weather conditions; (2) use explainable AI to interpret the model by evaluating the contribution of predictors to the wildfire risk. The modeling framework is shown in Figure 1. This paper is structured as follows: Section 2 describes the data and methods used, Section 3 highlights the key results, and next steps are described in Section 4.

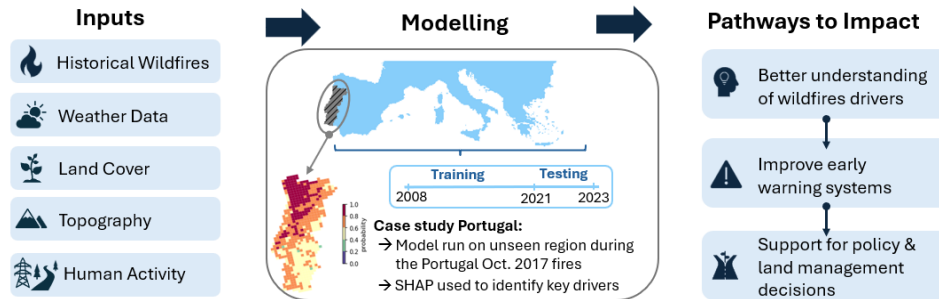


Figure 1: Wildfire risk modeling framework.

2 Data and methods

Data selection & processing

Fires recorded between 2008 and 2023 during the fire season (defined as June to October) with burned areas exceeding 50 ha were obtained from EFFIS (European Forest Fire Information System) and used as the predictand. Key predictors were selected based on a thorough literature review and include human infrastructure and land use (Bountzouklis et al., 2023; Oliveira et al., 2012; Zhai et al., 2023), as well as topographical and climate data (Polash, 2022; Dorph et al., 2022; Milanović et al., 2021). In total, 25 variables were aggregated into a common 12×12 km grid, providing a balance between computational efficiency and the ability to capture regional variation in the data. A detailed list of the variables is provided in Table A of the Appendix.

Generating the ML dataset

The ML dataset includes all of the Southern Europe domain, with the exception of Portugal, which is later applied as the case study of the fine-tuned model. The years 2008 to 2020 were used for training, and the years 2021 to 2023 for testing. Unique combinations of location and date, distinct from the historical record of fires, were selected to generate a balanced dataset consisting of 50% fire and 50% no fire data. A value of 1 (0) was assigned to the fire (no fire) data, and the corresponding 25 predictors were appended. Similar to Oliveira et al. (2012), pairs of predictors with a Pearson's correlation coefficient greater than 0.75 were identified, and one of the correlated predictors was removed. This process dropped five features to reduce the multicollinearity. Finally, twenty predictors were standardized by subtracting the mean and dividing by the standard deviation.

Training and testing the ML models

Four models were trained using the Scikit-learn Python library: Support Vector Machine (SVM), Random Forest (RF), k-Nearest Neighbors (kNN), and Logistic Regression (LR). The models were selected based on their interpretability, robustness, and performance. They have commonly been used in classification tasks in climate modelling studies (Polash, 2022; Milanović et al., 2021; Oliveira et al., 2012). The model hyperparameter tuning was performed on the training data with a 5-fold cross-validation using the F1 score. The final model selection was performed on the testing data and evaluated using three metrics: the F1 score, logarithmic loss, and Area Under the Curve (AUC).

Assessing the scalability of the model

In October 2017, devastating wildfires were recorded in Portugal, leading to 51 fatalities (Viegas et al., 2019). This month was selected to assess the performance and scalability of the model to unseen data. The Fire Weather Index (FWI), currently adopted by the Copernicus Emergency Management Service to quantify wildfire danger in Europe, was employed as the baseline to compare with the ML

results. The FWI was calculated from the daily temperature, precipitation, wind speed, and relative humidity following the method of Van Wagner (1974) .

3 Results

3.1 Model evaluation & selection

The four fine-tuned models with the optimal hyperparameters were evaluated on the testing data, shown in Table A1. The RF and SVM present the best-performing models, with the same AUC value of 0.91 (Figure A1 a), the same logarithmic loss of 0.38, and F1 scores of 0.83 and 0.84, respectively (Table A1). The RF model was selected for further analysis in this study because of its superior scalability and low computational cost compared to SVM (Zhang et al., 2023). Maps of the wildfire risk were generated for each day of the testing years (2021 - 2023) and were found to align well with the locations of the historical fires in the same period (Figure A1 b-c).

3.2 Model explainability using SHAP

The SHapley Additive exPlanations (SHAP) method can be used to assess the contribution of each feature to an individual prediction (Lundberg and Lee, 2017). The SHAP analysis was performed on the testing data to evaluate how the fine-tuned RF model quantified the fire risk in Southern Europe. The mean absolute SHAP values show the influence of the top-twelve-ranked features on the model's prediction (Figure 2.a). And the layered violin plot provides corresponding interpretability by indicating the relationship between the feature value, including both magnitude and positive and negative directions, and its impact on the model's prediction (Figure 2.b).

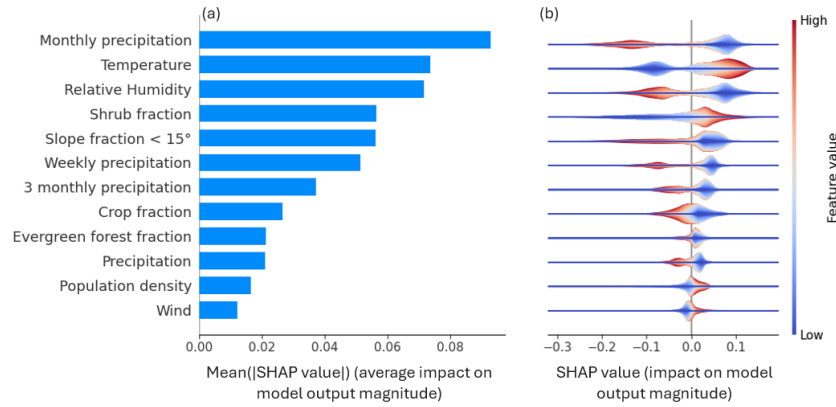


Figure 2: (a) Mean absolute SHAP values of the 12 most important predictors from the RF model on the testing dataset (b) SHAP value distributions for the same predictors, with colors showing the impact of the feature values on the output (red = high, blue = low).

The top three features are the monthly precipitation, temperature and relative humidity. Low monthly precipitation, low relative humidity, and high temperatures have a significant positive impact on the model's output, increasing the likelihood of wildfire occurrence (higher wildfire risk). A wide range of studies has confirmed these patterns, showing the impact of hot and dry periods on wildfire activity (Dong et al., 2021; Lanet et al., 2024). The next two most important features are the shrub fraction and the fraction of slope below 15°. The SHAP values also indicate that a higher shrubland fraction is positively related to the prediction of the occurrence of the wildfire. This provides evidence of the model's explainability, as shrublands are among the land cover types most prone to fire and can spread the fire very rapidly (Stavi, 2019). In contrast, high fractions of slope below 15° negatively affect the predicted occurrence of wildfire, which was explained in Pereira et al. (2016) that wildfires have been found to spread faster where the slope is steeper, especially when the slope is uphill.

3.3 Case study: the October 2017 wildfires in Portugal

The RF model was run for each day in October 2017 in Portugal. Figure 3 shows the number of recorded wildfires greater than 50 ha (panel a), the daily wildfire probability from the ML model (panel b) and FWI values (panel c) on the 15th of October. The date was selected as it was one of the days with the most recorded fires in the data. Most of the fires were recorded in the northwest of Portugal. On that day, all of Portugal was classified as being in a very high (FWI > 38) or extreme (FWI > 50) fire danger area, according to the EFFIS categories. The FWI is, however, not able to specifically depict the areas where wildfires were recorded. In contrast, the region with the higher wildfire probability (0.8-1.0) calculated by the RF model aligns closely with the recorded fires on the same day.

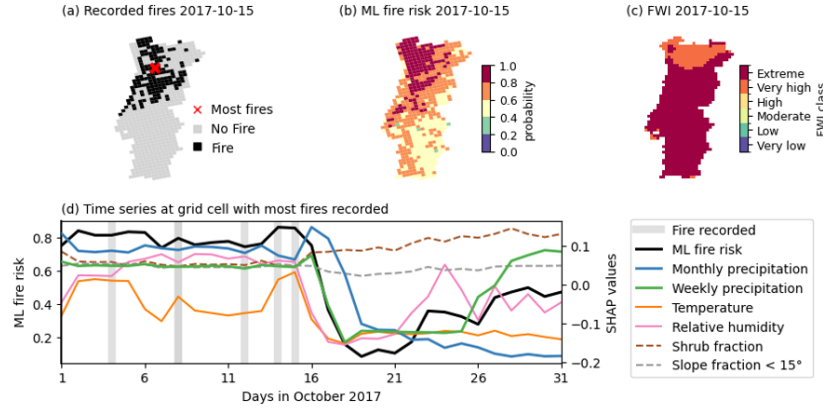


Figure 3: (a) Historical number of fires recorded per grid cell (b) Wildfire probability from the RF model (c) Fire Weather Index (FWI) on 15th of October 2017 (d) Time series of the ML probability at the grid cell with the most fires recorded in October 2017 (red cross in (a)) as well as the SHAP values for the top 6 features.

The time series of the RF model probability at one specific grid cell (grid cell with the most recorded fires - red cross in Figure 3.a), along with the SHAP values of the top 6 features obtained from a SHAP analysis on the dataset of Portugal in October 2017 are shown in Figure 3.d. The recorded fires occurred in the first half of the month, coinciding with high ML probabilities (> 0.7). From the 16th of October onward, the predicted probabilities of wildfire occurrence drop, and no additional fires were recorded. Monthly precipitation presents the highest mean absolute SHAP value, followed by relative humidity, weekly precipitation, and temperature, all of which drop in the middle of the month, in agreement with the drop in the predicted probability. The shrub fraction and the slope fraction below 15° also contributed to the model output, helping to explain why certain regions are more vulnerable to wildfire risk: The areas that experienced fires correspond to regions with mountainous terrain and a high proportion of shrubs, underscoring their susceptibility to wildfire (Figure A2).

4 Conclusion

In this study, four ML models were trained on a combined dataset of weather, human activity, topography, and land cover, and wildfire records in Southern Europe (excluding Portugal). The RF model (AUC = 0.91, F1 score = 0.85) was found to identify areas at high risk of wildfires more accurately than the FWI during the devastating October 2017 fires in Portugal, a region unseen by the model. The results highlight the scalability of the model and the added value of including non-weather variables. Using explainable AI, the most influential factors behind wildfire risk were identified, improving both the interpretability and trustworthiness of the model. This information can enable regional authorities in targeting zones with elevated fire probability and selecting relevant adaptation measures. Although demonstrated in one case, the framework is adaptable to other regions and time periods. Ongoing work focuses on assessing how wildfire risk evolves under different Shared Socioeconomic Pathways, incorporating novel datasets that not only represent the future climate but also its interaction with different land uses (Asselin et al., 2024).

Data Availability

Fire data was obtained from the European Forest Fire Information System (EFFIS) (Forest Fire Emergency Copernicus, 2025). The land cover data was obtained from the CORINE Land Cover inventory (European Environment Agency, 2020). The ERA5-Land variables were retrieved from the Climate Data Store (CDS) of Copernicus, available at (Muñoz Sabater, 2019). Topographical data was accessed through the Eurostat digital elevation model of Europe (GISCO, 2016). Power lines and road network data were extracted from (Arderne et al., 2020) and (Meijer et al., 2018), respectively. The population density was gathered from (Tatem, 2017). All datasets were processed using Python, standard geospatial libraries, and GIS software (QGIS). Upon publication of this paper, a GitHub repository will be made public with all code available.

References

- C. Arderne, C. Zorn, C. Nicolas, and E. E. Koks. Predictive mapping of the global power system using open data. *Scientific Data*, 7(1):19, 2020. ISSN 2052-4463. doi: 10.1038/s41597-019-0347-4. URL <https://doi.org/10.1038/s41597-019-0347-4>.
- O. Asselin, M. Leduc, D. Paquin, N. de Noblet-Ducoudré, D. Rechid, and R. Ludwig. Blue in green: forestation turns blue water green, mitigating heat at the expense of water availability. *Environmental Research Letters*, 19(11):114003, 2024.
- C. Bountzouklis, D. M. Fox, and E. Di Bernardino. Predicting wildfire ignition causes in southern france using explainable artificial intelligence (xai) methods. *Environmental Research Letters*, 18, 4 2023. ISSN 17489326. doi: 10.1088/1748-9326/acc8ee.
- J. Castel-Clavera, F. Pimont, T. Opitz, J. Ruffault, R. Barbero, D. Allard, and J.-L. Dupuy. A comparative analysis of fire-weather indices for enhanced fire activity prediction with probabilistic approaches. *Agricultural and Forest Meteorology*, 361:110315, 2025. ISSN 0168-1923. doi: <https://doi.org/10.1016/j.agrformet.2024.110315>. URL <https://www.sciencedirect.com/science/article/pii/S0168192324004283>.
- L. Dong, L. R. Leung, Y. Qian, Y. Zou, F. Song, and X. Chen. Meteorological environments associated with california wildfires and their potential roles in wildfire changes during 1984–2017. *Journal of Geophysical Research: Atmospheres*, 126(5):e2020JD033180, 2021. doi: <https://doi.org/10.1029/2020JD033180>. URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JD033180>. e2020JD033180 2020JD033180.
- A. Dorph, E. Marshall, K. A. Parkins, and T. D. Penman. Modelling ignition probability for human- and lightning-caused wildfires in victoria, australia. *Natural Hazards and Earth System Sciences*, 22:3487–3499, 10 2022. ISSN 16849981. doi: 10.5194/nhess-22-3487-2022.
- S. Erni, X. Wang, T. Swystun, S. W. Taylor, M.-A. Parisien, F.-N. Robinne, B. Eddy, J. Oliver, B. Armitage, and M. D. Flannigan. Mapping wildfire hazard, vulnerability, and risk to canadian communities. *International Journal of Disaster Risk Reduction*, 101:104221, 2024. ISSN 2212-4209. doi: <https://doi.org/10.1016/j.ijdr.2023.104221>. URL <https://www.sciencedirect.com/science/article/pii/S221242092300701X>.
- European Environment Agency. *CORINE Land Cover 2018 (raster 100 m), Europe, 6-yearly*. Copernicus Programme, European Environment Agency, 2020. doi: 10.2909/960998c1-1870-4e82-8051-6485205ebbac.
- European Environment Agency. *European climate risk assessment*. Publications Office of the European Union, 2024. doi: doi/10.2800/8671471.
- Forest Fire Emergency Copernicus. European Forest Fire Information System (EFFIS). <https://forest-fire.emergency.copernicus.eu/>, 2025. Accessed: 2025-08-06.
- GISCO. *EU-DEM: Digital Elevation Model*. Copernicus Programme, European Environment Agency, 2016. URL <https://ec.europa.eu/eurostat/web/gisco/geodata/digital-elevation-model/eu-dem>.

- M. C. Iban and O. Aksu. Shap-driven explainable artificial intelligence framework for wildfire susceptibility mapping using modis active fire pixels: An in-depth interpretation of contributing factors in izmir, türkiye. *Remote Sensing*, 16(15), 2024. ISSN 2072-4292. doi: 10.3390/rs16152842. URL <https://www.mdpi.com/2072-4292/16/15/2842>.
- M. Lanet, L. Li, A. Ehret, S. Turquety, and H. Le Treut. Attribution of summer 2022 extreme wildfire season in southwest france to anthropogenic climate change. *npj Climate and Atmospheric Science*, 7(1):267, 2024. ISSN 2397-3722. doi: 10.1038/s41612-024-00821-z. URL <https://doi.org/10.1038/s41612-024-00821-z>.
- S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
- J. R. Meijer, M. A. J. Huijbregts, K. C. G. J. Schotten, and A. M. Schipper. Global patterns of current and future road infrastructure. *Environmental Research Letters*, 13(6):064006, 2018. doi: 10.1088/1748-9326/aabd42. URL <https://doi.org/10.1088/1748-9326/aabd42>.
- S. Milanović, S. D. Milanović, N. Marković, D. Pamučar, L. Gigović, and P. Kostić. Forest fire probability mapping in eastern serbia: Logistic regression versus random forest method. *Forests*, 12:1–17, 1 2021. ISSN 19994907. doi: 10.3390/f12010005.
- J. Muñoz Sabater. *ERA5-Land hourly data from 1950 to present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2019. doi: 10.24381/cds.e2161bac.
- S. Oliveira, F. Oehler, J. San-Miguel-Ayanz, A. Camia, and J. M. Pereira. Modeling spatial patterns of fire occurrence in mediterranean europe using multiple regression and random forest. *Forest Ecology and Management*, 275:117–129, 7 2012. ISSN 03781127. doi: 10.1016/j.foreco.2012.03.003.
- P. Pereira, A. Cerdà, A. J. Lopez, L. M. Zavala, J. Mataix-Solera, V. Arcenegui, I. Misiune, S. Keesstra, and A. Novara. Short-term vegetation recovery after a grassland fire in lithuania: The effects of fire severity, slope position and aspect. *Land Degradation & Development*, 27(5):1523–1534, 2016.
- B. Polash. Modis-firms and ground-truthing-based wildfire likelihood mapping of sikkim himalaya using machine learning algorithms. *Natural Hazards*, 110:899–935, 1 2022. ISSN 15730840. doi: 10.1007/s11069-021-04973-6.
- M. Shafapourtehrany. Geospatial wildfire risk assessment from social, infrastructural and environmental perspectives: A case study in queensland australia. *Fire*, 6(1), 2023. ISSN 2571-6255. doi: 10.3390/fire6010022. URL <https://www.mdpi.com/2571-6255/6/1/22>.
- J. J. Shaples. A note on fire weather indices. *International Journal of Wildland Fire*, 31:728–734, 2022. doi: 10.1071/WF21134. URL <https://doi-org.proxy.findit.cvt.dk/10.1071/WF21134>.
- I. Stavi. Wildfires in grasslands and shrublands: A review of impacts on vegetation, soil, hydrology, and geomorphology. *Water*, 11(5):1042, 2019.
- A. J. Tatem. Worldpop, open data for spatial demography, 2017. ISSN 2052-4463. URL <https://doi.org/10.1038/sdata.2017.4>.
- C. E. Van Wagner. *Structure of the Canadian forest fire weather index*, volume 1333. Environment Canada, Forestry Service Ottawa, ON, Canada, 1974.
- D. X. Viegas, M. F. Almeida, L. M. Ribeiro, J. Raposo, M. T. Viegas, R. Oliveira, D. Alves, C. Pinto, A. Rodrigues, C. Ribeiro, S. Lopes, H. Jorge, and C. X. Viegas. Análise dos incêndios florestais ocorridos a 15 de outubro de 2017. Technical report, Centro de Estudos sobre Incêndios Florestais, ADAI/LAETA, Departamento de Engenharia Mecânica, Faculdade de Ciências e Tecnologia, Universidade de Coimbra, January 2019.
- J. Zhai, Z. Ning, R. Dahal, and S. Yang. Wildfire susceptibility of land use and topographic features in the western united states: Implications for the landscape management. *Forests*, 14(4), 2023. ISSN 1999-4907. doi: 10.3390/f14040807. URL <https://www.mdpi.com/1999-4907/14/4/807>.

B. Zhang, F. K. Abu Salem, M. J. Hayes, K. H. Smith, T. Tadesse, and B. D. Wardlow. Explainable machine learning for the prediction and assessment of complex drought impacts. *Science of The Total Environment*, 898:165509, Nov. 2023. ISSN 1879-1026. doi: 10.1016/j.scitotenv.2023.165509. URL <https://doi.org/10.1016/j.scitotenv.2023.165509>.

A Appendix

Model	Training		Testing		
	Hyperparameters (selected ones for testing in blue)	Best score from CV (F1)	AUC	F1	Logloss
SVM	C: [0.1, 1 , 10, 100, 1000] gamma: [1, 0.1 , 0.01, 0.001, 0.0001] kernel: ['rbf']	0.87	0.91	0.84	0.38
RF	n_estimators: [50, 100, 150 , 200] max_depth: [5, 10, 15, 20, 25 , 30]	0.88	0.91	0.83	0.38
kNN	n_neighbors: [5, 10, 15 , 20, 25, 30, 35]	0.86	0.90	0.83	0.72
LR	C: [0.001 , 0.01, 0.1, 1, 10, 100, 1000] solver: ['newton-cg', 'lbfgs', ' liblinear ', 'sag', 'saga']	0.84	0.88	0.81	0.44

Table A1: Comparison of machine learning models (SVM, RF, kNN, and LR) with the list of hyperparameters used for model tuning (with selected values highlighted in blue), best cross-validation F1 scores, and testing performance metrics (AUC, F1, and Logloss)

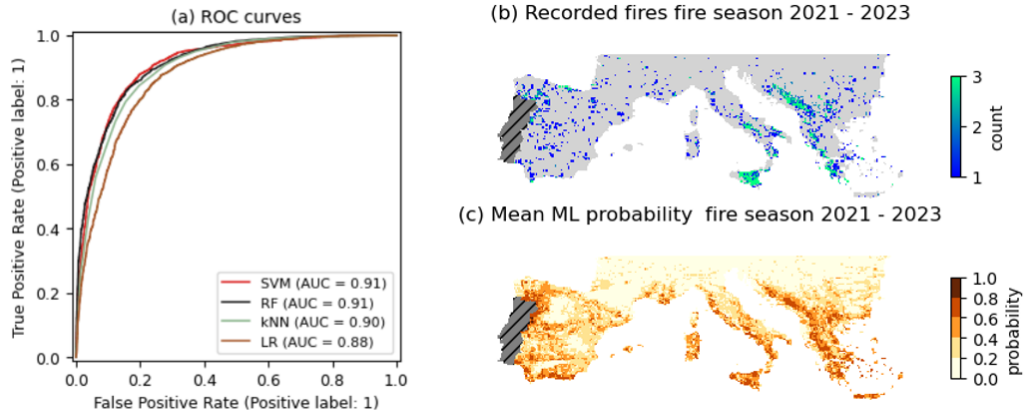


Figure A1: (a) ROC curves for the 4 models on testing data. (b) Historical recorded fires (c) Mean ML probability from RF model during the fire season between 2021 - 2023.

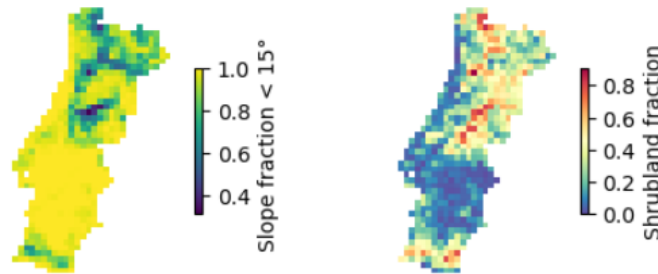


Figure A2: Slope fraction < 15° and shrubland fraction

Table A2: Categories and variables used as predictors in the model

Dataset Type	Variable		Description	Source
Climate Data	Temperature	°C	All variables represent the midday value, obtained at local noon. Derived variables (e.g., 7-day moving averages or weekly sums) are based on aggregate noon data up to the reference day. The value of the nearest ERA5-Land grid has been used	ERA5-Land, 0.1° hourly data from 1959 to present (Muñoz Sabater, 2019)
	Wind Speed	m/s		
	Relative Humidity	%		
	Daily Precipitation	m		
	Weekly precipitation Sum	m		
	Monthly precipitation Sum	m		
	3 Months precipitation sum	m		
Topographic	Mean Elevation	m	Average elevation above the sea level for the grid cell	Digital Elevation model over Europe, 30 m (GISCO, 2016)
	Slope >30°	-	Proportion (0–1) of the grid cell area where the slope exceeds 30°	
	15° < Slope < 30°	-	Proportion (0–1) of the grid cell area where the slope is between 15° and 30°	
	Slope < 15°	-	Proportion (0–1) of the grid cell area where the slope is below 15°	
	Aspect N	-	Proportion (0–1) of the grid cell area covered by each aspect direction: North, East, South, and West	
	Aspect E	-		
	Aspect S	-		
Land Cover	Grassland	-	Proportion (0–1) of the grid cell area covered by each different land cover type. The classes have been obtained following the categories from Zhai et al. (2023)	CORINE Land Cover 100m (European Environment Agency, 2020)
	Broadleaf Forest	-		
	Evergreen Forest	-		
	Cropland	-		
	Shrubland	-		
	Urban	-		
	Water	-		
Human Infrastructure	Road Density	m/km ²	Ratio between the sum of total available road network length, including highways, primary, secondary and tertiary roads and cell area	Global Roads Inventory Project (GRIP) dataset (Meijer et al. 2018)
	Power Line Density	m/km ²	Ratio between the total sum of electric lines within a cell and its area	Global Power System Mapping dataset (Arderne et al. 2020)
	Population Density	per/ km ²	Ratio of inhabitants per grid cell	WorldPop Global Population dataset (Tatem, 2017)