Capitalizing the Predictive Potential of Machine Learning to Detect Various Fire Types Using NASA's MODIS Satellite Data for the Mediterranean Basin

Nima Kamali Lassem
Department of Information Systems and Technologies, Bilkent University, Türkiye

Obai Mohamed Hisham Abdelmohsen Gaafar
Department of Computer Science, University of Milan, Italy

Seyid Amjad Ali
Department of Information Systems and Technologies, Bilkent University, Türkiye

ABSTRACT

This study investigates the realm of machine learning for the classification of different fire types using NASA's FIRMS MODIS satellite data for the Mediterranean basin. Concentrating on the Mediterranean basin and utilizing data spanning from 2019 to 2021 for model training, XGBoost and Random Forest models were subsequently validated for the 2022 data. The findings distinctly illustrate XGBoost's superior predictive precision as compared to Random Forest by showcasing an impressive overall F1 score surpassing 95% and 84% macro F1 score across various fire types. This study emphasizes the prospect of machine learning to improve worldwide wildfire monitoring and response by providing exact, real-time fire type forecasts.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Machine learning approaches**; • **Classification and regression trees**;

KEYWORDS

Wildfire prediction, MODIS, Mediterranean basin, XGBoost, Random Forest

ACM Reference Format:
Nima Kamali Lassem, Obai Mohamed Hisham Abdelmohsen Gaafar, and Seyid Amjad Ali. 2023. Capitalizing the Predictive Potential of Machine Learning to Detect Various Fire Types Using NASA's MODIS Satellite Data for the Mediterranean Basin. In *2023 The 7th International Conference on Advances in Artificial Intelligence (ICAAI) (ICAAI 2023), October 13–15, 2023, Istanbul, Turkiye.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3633598.3633603

1 INTRODUCTION

In the face of intensifying forest fire occurrences, the accurate monitoring of these events has become critical. While forest fires are integral to our ecosystems, their growing harmfulness poses serious threats to infrastructure, human settlements, and biodiversity. Beyond immediate devastation, wildfires can disrupt water bodies and have far-reaching consequences on the environment, as noted in 'Wildfire's Impact on Our Environment' [1]. Furthermore, the hazardous pollutants in wildfire smoke, as underlined by WHO, have a direct influence on public health [2]. NASA's Doug Morton anticipates a rise in forest fires across the US by 2050 [3], emphasizing the need for effective fire monitoring. Recently, due to the aforementioned reasons, there has been a major upsurge in the number of studies that are being conducted by the research community to highlight this serious issue. Authors in [4, 5] have provided a detailed review regarding forest fires and some machine learning based algorithms that can be used for their detection. Similarly, Gözde et al., [6] used a dataset[1] that consists of fires in a national park in northern Portugal between January 2000 and December 2003 to compare the performance of various machine learning algorithms.

Amid severe challenges posed by forest fires, systems like NASA's Fire Information for Resource Management System (FIRMS) have revolutionized disaster management. Yet, accurately predicting fire types remains crucial for assessing threats and aiding rescue services. This paper highlights the importance of predicting fire types observed by FIRMS MODIS satellites by using novel machine-learning methods. By addressing gaps in fire type prediction, we seek to enhance global disaster response. Through real-world scenarios, including volcano-related and non-vegetation fires, we showcase machine learning's pivotal role in fire type prediction. Notably, to the best of our knowledge, this dataset has never been used to perform multiclass classification of different fire types, therefore our work adds substantial novelty to this field of study.

2 REMOTE SENSING AND FIRE MONITORING

The Earth Observing System (EOS), a comprehensive program that uses a range of methodologies to study the events on our planet, is a component of NASA's devotion to monitoring the planet. The extraterrestrial Firewatch system, exemplified by the Moderate Resolution Imaging Spectroradiometer (MODIS), is crucial to this effort. As articulated by the Yale Center for Environmental Law & Policy (2021), MODIS stands as an expansive program facilitated by sensors onboard two satellites, collectively ensuring complete daily coverage of the Earth. This comprehensive coverage is achieved by leveraging an array of resolutions—spectral, spatial, and temporal—thus enabling a nuanced understanding of the environment. Notably, the MODIS sensor operates on both the Terra and Aqua satellites, affording the availability of imagery in both morning (Terra) and afternoon (Aqua) timeslots for any specific location. Even during nighttime, data remains accessible in the thermal range of the electromagnetic spectrum [7].

[1] https://archive.ics.uci.edu/dataset/162/forest+fires

ICAAI 2023, October 13–15, 2023, Istanbul, Turkiye

ACM ISBN 979-8-4007-0898-5/23/10.
https://doi.org/10.1145/3633598.3633603

Table 1: NASA FIRMS MCD14DL-NRT Attributes [8].

Attribute	Short Description
Latitude	Latitude
Longitude	Longitude
Brightness	Brightness temperature 21 (Kelvin)
Scan	Along Scan pixel size
Track	Along Track pixel size
Acq_Date	Acquisition Date
Acq_Time	Acquisition Time
Satellite	Satellite
Confidence	Confidence (0-100%)
Version	Version (Collection and source)
Bright_T31	Brightness temperature 31 (Kelvin)
FRP	Fire Radiative Power (MW - megawatts)
Type*	Inferred hot spot type: 0 = presumed vegetation fire 1 = active volcano 2 = other static land source 3 = offshore
DayNight	Day or Night

Figure 1: Map of the Mediterranean Basin as defined in this research.

Accessing the wealth of fire-related data collected by the MODIS system is facilitated through NASA's FIRMS API. This interface provides a conduit for acquiring fire data in the form of comma-separated values (CSV), including an array of features encompassing crucial information. Some of the attributes included can be seen in Table 1.

Unfortunately, the attribute 'Type', as seen in the table is unavailable through the Near Real Time (NRT) API service offered by NASA FIRMS. This attribute, however, is included in the data from the previous years. Researchers interested in accessing and utilizing this observed dataset[2] for different years can do so effortlessly. It is worth mentioning here that as of late August 2023, users are now required to create requests on the NASA FIRMS to access the data used for this study.

2.1 Mediterranean Basin

Upon an initial survey of the global distribution of fires on the world map, one promptly observes the prevalence of numerous fire outbreaks spanning the planet, thus yielding extensive data for predictive purposes. Researchers have conducted studies of wildfires and human behavior in regions such as Australia and the USA. Regions in Europe have received less attention, despite facing the same issues, according to the study "Island Vulnerability and Resilience

[2]https://firms.modaps.eosdis.nasa.gov/country

Table 2: gives provides descriptive statistics on the Mediterranean Basin dataset we filtered earlier from the NASA FIMRS MODIS dataset for the years 2019 to 2021.

	latitude	longitude	scan	track	acq_time	bright_t31	frp
Min	27.647	-18.133	1.000	1.000	0.000	265.100	0.000
Mean	38.518	17.990	1.555	1.193	1182.082	300.348	54.061
Std	3.544	16.076	0.783	0.240	509.545	11.042	179.534
Q1	36.690	5.460	1.000	1.000	953.000	292.500	9.100
Median	37.484	17.150	1.200	1.100	1113.000	299.000	16.700
Q3	41.348	34.987	1.700	1.300	1257.000	308.100	38.700
Max	46.760	43.580	4.800	2.000	2359.000	400.100	11275.800

to Wildfires: A Case Study of Corsica" [9]. Consequently, a strategic decision was made to confine our research to a specific geographic area—the Mediterranean basin. This region exhibits a multitude of fire incidents attributed to its diverse biomes and the consistent fluctuations in temperature and overall climate. It's worth noting that the Mediterranean basin holds various definitions, and for the purpose of this study, we adopted the description provided on Wikipedia[3]. Subsequently, we proceeded to cartographically represent this area by using MapBox.

Considering the worrisome shifts in climate patterns observed in recent times, such as the escalation of global temperatures and an upsurge in severe wildfires, we acknowledged the necessity to orient our research toward contemporary climate information. Given the swiftly evolving nature of these developments, we confined our dataset to encompass solely the data from the preceding four years, excluding the ongoing year of 2023. The time span of 2019 to 2021 served as the foundational training dataset for the construction of our predictive framework, enabling it to derive insights from the most up-to-date trends in temperature oscillations. Subsequently, the complete data from the year 2022 was employed to evaluate and validate the models' performance in generating precise projections concerning forthcoming wildfire risks based on the latest climate tendencies.

2.2 Descriptive Statistics

As the attributes 'brightness' and 'bright-t31' captured the identical variable using distinct methodologies and presented data in uniform units (Kelvin), we opted to evaluate the correlation between these attributes within the dataset to prevent the use of redundant features during model training. The assessment of correlation was executed using the 'Pearson' technique, and the corresponding outcomes are provided in Figure 2 via a heatmap. The visual representation illustrates a moderate positive correlation between the two attributes. Additionally, it becomes apparent that the attribute 'brightness' displays a more pronounced correlation with the 'frp' attribute (Fire Radiation Power, mw). Consequently, to eliminate redundancy, the 'brightness' attribute is excluded from consideration. Furthermore, attributes such as 'confidence', 'acq_date', 'acq_time', 'satellite', 'version', and 'daynight' were also omitted from the dataset due to either missing values or their dependency on other columns, thus rendering them redundant.

[3]https://en.wikipedia.org/wiki/Mediterranean_Basin

The Pearson coefficient, also called the Pearson correlation coefficient, measures the strength of a linear relationship between two variables (X and Y) plotted on a scatter plot. With values ranging from -1 to +1, it gauges resemblance to a straight line: +1 signifies a perfect positive relationship, -1 indicates a perfect negative relationship, and 0 means no correlation between the variables [10].

Conversely, the attributes 'scan' and 'track' exhibit a significant correlation with a coefficient of 0.98. This observation prompted us to delve deeper into the possibility of redundancy between these attributes. Upon closer examination, however, it becomes evident that these attributes do not represent the same variables. According to information provided on the NASA Earthdata website, the 'scan' value signifies the spatial resolution in the East-West direction of the scan, while the 'track' value signifies the North-South spatial resolution of the scan. Notably, the pixel size is not uniformly 1 km across the scan track; it is larger than 1 km at the "Eastern" and "Western" edges of the scan, being 1 km only along the nadir, or the exact vertical from the satellite. Consequently, the reported values for 'scan' and 'track' accurately depict the genuine spatial resolution of the scanned pixel. This clarifies that the 'track' and 'scan' attributes encapsulate distinct facets of spatial resolution and do not possess redundancy. In the next section, we provide details about model training.

3 MACHINE LEARNING SOLUTIONS

Random forest is an extremely popular decision trees based machine learning algorithm trademarked by Leo Breiman [11] which it known for its ease of use and flexibility and can handles both classification and regression problems. Given that our aim involves executing multiclass classification across our four fire types (0 = presumed vegetation fire, 1 = active volcano, 2 = other static land sources, and 3 = offshore), this algorithm aligns well with our objectives. Another machine learning algorithm that also harnesses the ability of decision trees is eXtreme Gradient Boosting (XGBoost) algorithm [12]. It leverages distributed gradient-boosted decision trees (GBDT) and finds common applications in solving regression, classification, and ranking problems.

The confusion matrix map was used to evaluate the performance of the employed model using values of TP, TN, FP, and FN metrics. Accuracy, recall, precision, and F1 scores are computed according to eqs. 1–7. All the results are tabulated in Tables 3 to 5. In this study, the F1 Score is assessed using three methods: Micro, Macro, and Weighted. The micro method involves calculating metrics on

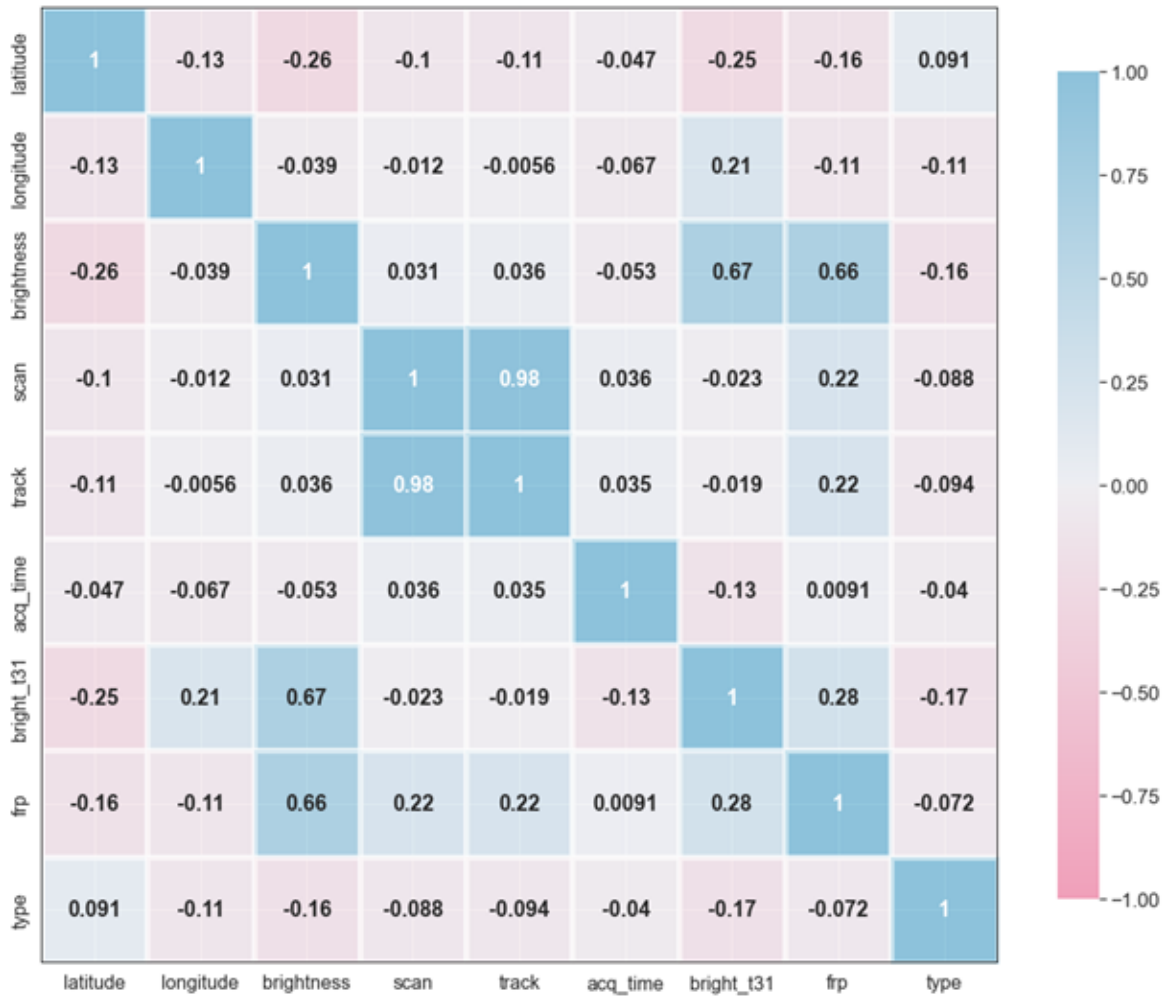

Figure 2: Correlation matrix for the Mediterranean Basin dataset from 2019 - 2021.

a global scale by tallying total TP, FN, and FP. The macro method calculates metrics for each label and determines their unweighted mean. Meanwhile, the weighted method computes metrics for each label and derives their weighted average based on support values [13].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

$$Micro\ F1\ Score = \frac{TP}{TP + 0.5\,(FP + FN)} \tag{5}$$

$$Macro\ F1\ Score = \frac{\sum_{i=1}^{number\ of\ classes} F1\ Score_i}{number\ of\ classes} \tag{6}$$

$$Weighted\ F1\ score = \sum_{i=1}^{number\ of\ classes} w_i F1\ Score_i \tag{7}$$

Embarking with Random Forest, the training process was executed using Python within a Google Colab[4] session, utilizing the training dataset (comprising 98% of the original dataset's size). Unfortunately, the validation dataset (constituting 2% of the original dataset's size) could not be used for Random Forest in order to implement early stopping, as the library does not support this feature.

The results show high accuracy of the model with a global (micro) and weighted F1 score of more than 94%. These metrics show the model remains accurate when considering the dataset as a whole. However, the F1 score of above 67% across the classes (Macro) is not very exciting. Even though, it is a reasonable good result, we can see that the model has a small problem with recall at 57% (Table 3).

As you can see in Table 3, the Random Forest model seems to be struggling with class 3. This is due to the imbalanced nature of

[4]https://colab.google/

Table 3: XGBoost and Random Forest various performance score.

Metrics	Random Forest	XGBoost
Accuracy	0.946	0.956
Micro F1 Score	0.946	0.956
Macro F1 Score	0.667	0.771
Weighted F1 Score	0.942	0.954
Precision	1.000	1.000
Recall	0.571	0.714
Class 0	0.969	0.974
Class 1	0.727	0.833
Class 2	0.801	0.842
Class 3	0.169	0.435

the dataset. Class 3 contributes to a much smaller proportion of the dataset than the other classes. Though model inaccuracies may seem minor, it's vital to remember our study's ultimate goal i.e., we can't afford to make errors as they could impact lives. Hence, we must prioritize improving our outcomes to the fullest extent possible. To address this, we consider the potential of gradient boosting as stated by Chamandeep et al. [14].

The results for the XGBoost model not only demonstrate significantly higher overall accuracy but also illustrate impressive performance across individual classes. The Macro F1 score surpasses 77%, indicating excellent accuracy across the classes. Furthermore, the F1 scores for each class, as depicted in Table 3, reveal nearly perfect accuracy for class 0, excellent accuracy for classes 1 and 2, and satisfactory accuracy for class 3, respectively. As observed in the work of Abdualgalil et al., [15], where they successfully employed XGBoost for COVID-19 infection prediction using clinical data, XGBoost continues to demonstrate its excellent performance in comparison to other machine learning algorithms. The test results clearly show that XGBoost greatly outperforms the Random Forest model, notably in class 3. It surpasses Random Forest's Macro F1 Score by an impressive 16%, illustrating outstanding overall performance on the Mediterranean Basin dataset.

4 CONCLUSION

This study demonstrates the remarkable efficacy of machine learning algorithms for the classification of different wildfire types by using satellite data to bolster global disaster response capabilities. Through the evaluation of XGBoost and Random Forest models on NASA FIRMS MODIS data for the Mediterranean basin, our results highlight the exceptional performance of XGBoost — achieving an overall F1 score exceeding 95% and a macro F1 score of 84% across various fire types. This research underscores the potential of machine learning techniques to extract valuable insights from Earth observation data, enabling a more proactive approach to fire monitoring and intervention. While this study showcases significant achievements, our future endeavors, including the experimentation with state-of-the-art technologies like BERT, to explore the realm of deep learning models to enhance predictive accuracy even further. By extending these accomplishments, our ongoing work involves the development of deep neural networks to improve performance on varied and imbalanced wildfire data. By improving real-time fire prediction, the goal is to enhance early warnings and protect at-risk communities globally, investigate the potential of more advanced and complex machine learning algorithms for proactive wildfire monitoring, henceforth, increasing life-saving efforts.

REFERENCES

[1] Jmendenhall, "Wildfire's impact on our environment," Utah Department of Environmental Quality, https://deq.utah.gov/communication/news/wildfires-impact-on-our-environment#:$\sim$:text$=$Wildfires%20can%20affect%20the%20physical,significant%20increase%20in%20stormwater%20runoff (accessed Aug. 28, 2023).
[2] "Wildfires," World Health Organization, https://www.who.int/health-topics/wildfires#tab$=$tab_1 (accessed Aug. 28, 2023).
[3] B. Dunbar, "Climate models project increase in U.S. wildfire risk," NASA, https://www.nasa.gov/topics/earth/features/climate-fire.html
[4] Ramez Alkhatib, Ramez Alkhatib, Wahib Sahwan, Anas Alkhatieb, Brigitta Schütt, "A Brief Review of Machine Learning Algorithms in Forest Fires Science," (2023), https://doi.org/10.3390/app13148275
[5] Piyush Jain, Sean C.P. Coogan, Sriram Ganapathi Subramanian, Mark Crowley, Steve Taylor, Mike D. Flannigan, "A review of machine learning applications in wildfire science and management," Environmental Reviews (2020) https://doi.org/10.1139/er-2020-0019
[6] Gözde BAYAT, Kazım YILDIZ, "Comparison of the Machine Learning Methods to Predict Wildfire Areas," Turkish Journal of Science & Technology (2022), Volume: 17 Issue: 2, pages 241 – 250. https://doi.org/10.55525/tjst.1063284
[7] "What is the landsat program?," Center for Earth Observation, https://yceo.yale.edu/faq-page (accessed Aug. 28, 2023).
[8] N. Earth Science Data Systems, "NASA MODIS Attributes, MCD14DL-NRT," NASA, https://www.earthdata.nasa.gov/learn/find-data/near-real-time/firms/mcd14dl-nrt#ed-firms-attributes (accessed Aug. 28, 2023).
[9] Sandra Vaiciulyte *et al.*, "Island vulnerability and resilience to wildfires: A case study of corsica," International Journal of Disaster Risk Reduction, https://doi.org/10.1016/j.ijdrr.2019.101272
[10] W. Kenton, "What is the Pearson coefficient? definition, benefits, and history," Investopedia, https://www.investopedia.com/terms/p/pearsoncoefficient.asp (accessed Aug. 28, 2023).
[11] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001), https://doi.org/10.1023/A:1010933404324
[12] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785
[13] "Sklearn.metrics.f1_score," scikit, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html (accessed Aug. 28, 2023).
[14] Bilal Abdualgalil, Sajimon Abraham, and Waleed M. Ismael, "COVID-19 Infection Prediction Using Efficient Machine Learning Techniques Based on Clinical Data," Journal of Advances in Information Technology, Vol. 13, No. 5, pp. 530-538, October 2022.