

Predicting forest fire : a new hybrid approach

Houda El Bouhissi^{1,*†}, Naima Illoul^{1,†}, Tetiana Hovorushchenko^{2,†} and Olga Pavlova^{2,†}

¹ University of Bejaia, Faculty of Exact Sciences, LIMED Laboratory, 06000, Bejaia, Algeria

² Khmelnytskyi National University, Instytut'ska str., 11, Khmelnytskyi, 29016, Ukraine

Abstract

In recent years, forest fires have become a significant problem that has attracted international attention, particularly in Algeria. These fires are one of the most pressing global security challenges of our time. Therefore, an assessment is made of the existing protective mechanisms in relation to these fires and how they protect the country's environment from such catastrophic destruction. There are many effective techniques in the field of forecasting. The most common of these are those based on artificial intelligence, including machine learning and related tools and models. These techniques have enabled the development of reliable and adaptive systems in various fields, particularly in the field of forecasting. We use these techniques to predict forest fires. The aim of this paper is to propose a new hybrid approach based on ML and bioinspired algorithms to predict forest fire. The experimental results show that the application of the bioinspired algorithm improves ML efficiency.

Keywords

Machine learning, Kaggle, datasets, forest fire prediction, logistic regression, Random Forest, Decision tree, GWO.

1. Introduction

Forest fires are among the world's most perilous natural disasters. They cause catastrophic losses to forest ecosystems and pose a serious threat to human safety and property. These conflagrations have the potential to cause substantial disruption to biodiversity, ecosystem functioning, and the environment of the affected area [1].

Algeria is one of the countries that most affected by these disasters on an annual basis [2]. The high risk associated with these events has led to significant concern among investors, who are questioning the effectiveness of protection measures against these powerful fires and their ability to safeguard the country's environment.

Despite the efforts made by the protection services to avoid forest fire, this problem remains a major risk for the country's environment and the safety of its population. The damage and danger left behind by these fires worry officials and associations in the country who are trying to find immediate solutions to put an end to this disaster by providing all the necessary equipment.

The experience of all these years proves that despite the immediate intervention of the protection services, it still generates a significant rate of damage; the country is still vulnerable by these forest fires every year in summer. For this reason, building a prediction system for forest fire seems an interesting solution to prevent the risks and reduce the damage.

MoMLeT-2025: 7th International Workshop on Modern Machine Learning Technologies, June, 14, 2025, Lviv-Shatsk, Ukraine

* Corresponding author.

† These authors contributed equally.

✉ Houda.elbouhissi@gmail.com (H. El Bouhissi); naimailloul2904@gmail.com (N. Illoul); t.tereshchenko@univer.km.ua (T. Tereshchenko); pavlovao@khnmu.edu.ua (O. Pavlova)

ORCID 0000-0003-3239-8255 (H. El Bouhissi); 0009-0001-3495-3710 (N. Illoul); 0000-0001-5748-2854 (T. Tereshchenko); 0000-0003-2905-0215 (O. Pavlova);



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Recently, the prediction domain has garnered significant attention, and research techniques are mostly based on artificial intelligence, in particular machine-learning (ML) which can be used to create reliable and robust prediction systems in various disciplines.

The aim of the present paper is to propose a new approach based on ML to predict forest fires in Algeria. This system can be used all over the world and is able to classify the possibility of forest fires into two categories (fire and non-fire) which is essential for the study.

The rest of the paper is organized as follows. Section 2 presents a comprehensive overview of the most important approaches associated with Fire Forest. In section 3, we describe our classification proposal in detail. Then in section 4, we present an empirical study of the proposed approach to assess its performance and efficiency. Finally, Section 5 concludes the paper and establishes the opportunity for future work.

2. Related Work

Forecasting forest fires through satellite imagery necessitates a systematic and thorough approach. This process involves extracting relevant data and developing predictive models. Within the realm of data analysis, some methodologies employ ML algorithms to enhance prediction accuracy.

Conversely, alternative approaches apply artificial neural networks, employing deep learning methodologies to enhance the precision of predictions. In this section, we present the most important contributions regarding forest fire prediction.

In their study, Abid and Izeboudjen (2020) [3] propose a predictive model, learned by decision tree (DT), for forest fire prediction in Algeria. The data used is gathered from two regions of North Algeria: Sidi-Bel-Abbes and Bejaia. The meteorological data analyzed in this study include three attributes that have a significant impact on fire behavior: temperature, relative humidity and wind speed. The obtained results underline the suitability of the DT approach for achieving the specific objective with high performance.

As stated by Mohajane et al. (2021) [4], the creation of forest fire maps is of crucial importance in ensuring the proper functioning of forest ecosystems and preserving their valuable benefits for human well-being. The authors combine historical fire occurrence data from the Fire Information for Resource Management System with Landsat 8 OLI satellite imagery and a digital elevation model from the Shuttle Radar Topography Mission. The meteorological and topographic variables are then derived and processed using the Spatial Analyst "surface" tools, resulting in the creation of high-resolution maps. These maps serve as effective decision-support for analyzing fire behavior and devising management strategies in any fire-prone region.

In a separate study, Preeti et al. [5] present a system that integrates comprehensive weather data from Kaggle. An exploratory analysis was conducted, followed by preprocessing aimed at eliminating noisy data and converting categorical variables into numerical ones, thereby enhancing the clarity of the dataset. Following this preliminary processing step, the identification of hotspots is based on the available meteorological data, and models are applied to predict the likelihood of a fire occurrence, with notifications being sent to the nearest station. The data presented in this proposal encompasses a range of meteorological parameters, including but not limited to rainfall, temperature, humidity, and wind. The process commences with the analysis of these data points, which results in the development of a regression model that utilizes weather-related variables. After data collection, the data undergoes a series of preprocessing steps to ensure that it adheres to the requisite standard format. The following step is to select the appropriate model for the dataset. The regression techniques employed for prediction purposes include Random Forest (RF), DT, Support Vector Regression (SVR), and Naive Bayes (NB).

Si et al. [6] proposed a novel approach using ML algorithms, specifically logistic regression (LR), to predict forest fire risk in the Lijiang region (China). The authors employ a methodology that seamlessly integrates spatial analysis, a binomial logistic model, and Kriging interpolation. This approach makes it possible to assess the influence of various factors on the study subject, such as topography (altitude, slope, and orientation), vegetation and weather conditions (precipitation,

temperature, wind, humidity). A model was developed based on empirical data collected from fire points and random non-fire points. This methodological approach made it possible to estimate the probability of fire occurrence based on the relationships between these two types of points. The best performing model was selected to predict fire risk.

Singh et al [7] propose a new method, namely parallel SVM, for reliable performance of forest fire prediction. The data used is collected from the Indian Meteorological Department, which consists of weather data from the Indian region. In the context of implementing the system, several frameworks have been employed. The Django framework is chosen for the user interface, while Apache Spark is used for massive data analysis. This type of solution can help very well with the detection of the fires before they destroy the whole forest and simplify the prediction of these forest fires.

Another interesting approach proposed by Soualah et al. [8] which aims to provide a comprehensive analysis of forest fire risk patterns around Djebel El Ouahch (Algeria). The study focuses on the integration of bioclimatic, fuel, geomorphological and human factors using advanced fuzzy logic and geographic information system techniques. Data from climate stations, satellite imagery and GIS were used to map the bioclimatic parameters, the land cover and the geomorphological features. The resulting forest fire risk prediction model was defuzzified to generate prediction maps that indicated different levels of vulnerability within the study area. According to the authors, the research has provided valuable insights into the risk in the Djebel El Ouahch and has served for developing fire management strategies.

In addition, Sathishkumar et al. [9] propose a new research paper, which looks at detection of fire/smoke from images using AI-based computer vision techniques such as Convolutional Neural Networks (CNNs). According to state-of-the-art, methods in image classification and other computer vision tasks, CNNs achieved better accuracy; however, training them can take a long time.

The authors Sharma and Khanal [10] employed various ML and DL algorithms for fire forest prediction with the focus on Feature importance analysis to construct the final fire prediction model. The data used was gathered from the south California Forestry Commission for the year 2023. The experiment results showed that the DT model achieved higher accuracy of fire prediction with 90.58% accuracy, followed by RF with 88% accuracy. The score for Artificial Neural Network (ANN) was low. Similarly, DT classification error was low (9.42%) and ANN high.

Furthermore, Zennir and Khallef [11] present a study, which aims to generate a forest fire probability map using Logistic Regression (LR) and SVM algorithms combined with GIS methods. The results showed that the accuracy of the fire probability map produced by the LR model was better (AUC = 0.845) than the accuracy of the map produced by the SVM model (AUC = 0.748). More than half of the forests in the study area had a very high/high probability of fire according to the probability maps.

Several papers covered describe in detail forest fire prediction methods that produce interesting results. The methods used in this study are based on techniques from ML, DL and data mining. ML has emerged as a widely involved technology, with fundamental algorithms such as RF, SVMs, and LR serving essential roles. The method is highly computationally efficient due to the application of various algorithms, thereby improving the performance and robustness of the system. On the other hand, conducting in depth research and attaining the desired outcomes can be quite challenging due to the complexities and the sheer volume of information involved.

Data mining is also presented as one of the methods commonly used in the field of forecasting using these different algorithms. The main advantage of this method is its ability to significantly reduce the time spent on finding a solution by employing various statistical techniques to analyze the data. However, a notable drawback is that it demands a considerable amount of computing resources.

DL is considered one of the most valuable approaches in the field, alongside machine learning. Its significance is clearly demonstrated through various algorithms and techniques, such as artificial neural networks (ANNs) and convolutional neural networks (CNNs), which lend them credibility and reliability. DL is distinguished by its ability to process unstructured data, such as images or

videos, with superior performance, but at the cost of significant computational resources and large amounts of data required for training.

Image mining is particularly well suited to the processing of satellite data. Although this approach is effective for detecting and classifying visual objects, it involves intensive data preprocessing, which can slow down the overall process and reduce system efficiency.

To take advantage of the previous works, we propose a new effective solution to predict forest fire based on a hybrid approach combining ML and bioinspired algorithm. The proposed methodology aims to provide a suitable and usable prediction system for forest fire to reduce fires risk. We use the most accurate ML algorithms (LR, RF and DT) regarding the related works. In addition, we apply the GWO a bioinspired algorithm to select the best features that fit with the experiment which optimize significantly the prediction process. Compared with the related solutions, our proposal improves the prediction accuracy.

In the newt section, we will describe our proposal in detail.

3. Proposed Methodology

Algeria is a distinctive ecological region within the biosphere. However, the lack of awareness regarding wildfire disasters and the absence of adequate funding to address these issues has contributed to the acceleration of desertification.

Furthermore, the Algerian forest, which is of significant ecological importance due to its biodiversity and its impact on the country's socioeconomic balance, requires protection. To ensure the success of development programs and to address environmental degradation, it is essential to identify computer-aided solutions and implement an integrated and participatory approach to reduce risks involved.

Therefore, there is a notable interest in forest fire forecasting systems. The methodology proposed in this study aims to predict forest fire occurrences. The system has been developed for the purpose of classifying forest regions as either "fire" or "non-fire" risk, with this classification being determined by meteorological conditions.

The system architecture as illustrated in figure 1, involved four distinct steps:

- The first step entitled "Data collection" concerns the data acquisition.
- Next, the second step "Data Processing" involves various phases for data preparation and refinement (for example data cleaning) to ensure data availability.
- The third step: "Feature selection" is about choosing the most important features for the next step. A bioinspired algorithm performs this step.
- The last step "Prediction" which is performed by a ML algorithm to obtain the result Fire/Not Fire.

3.1 Data Collection

The first phase to start the construction of our model is collecting data, which is a fundamental and critical part of our study because everything depends on it. We understand that the quality of the data collected is very important. To get more interesting and accurate results, it is important to collect reliable data.

This research is based on public data gathered from the Kaggle platform. Data is concerning two regions of northern Algeria called Bejaia and Sidi-Bel-Abbes to provide information about forest fires. The dataset contains distinct climate data and Fire Weather Index (FWI) components. These elements have been identified as contributing significantly to the initiation and spread of forest fires.

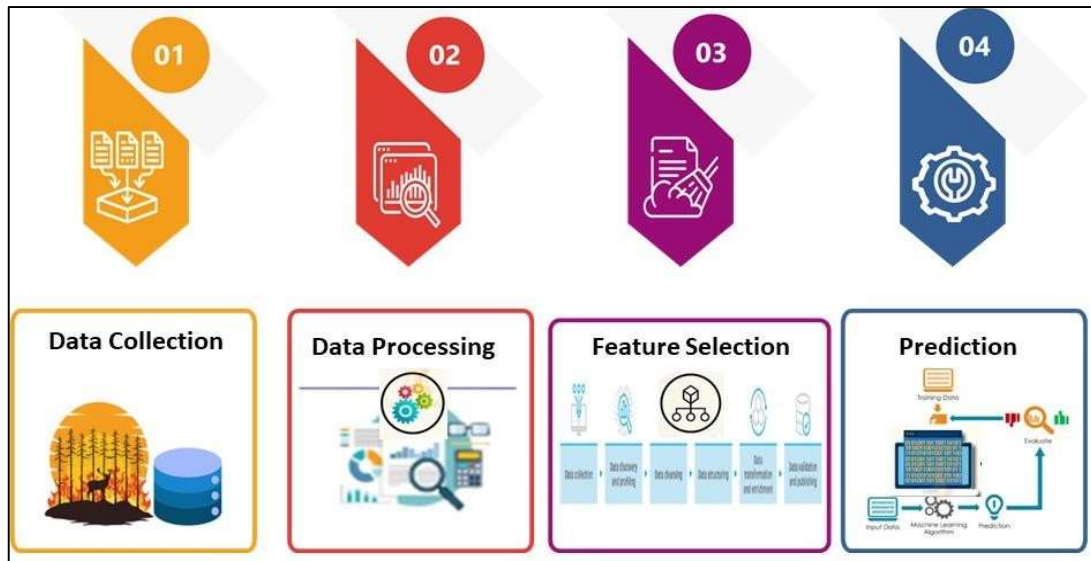


Figure 1: System architecture

3.2 Data processing

An exploratory data analysis was conducted to examine the relationship between variables. A heat map of correlation coefficients was drawn up, highlighting notable dependencies between the target variable (fire occurrence) and several input characteristics.

Among these factors, temperature, wind speed, FFMC, and FWI were found to be particularly significant (figure 2). This step not only facilitated the visualization of the distributions but also enabled the exploration of feature redundancy.

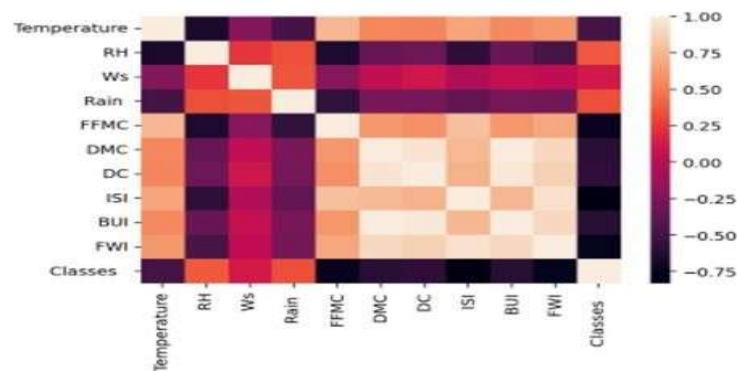


Figure 2: Heat map of feature (and outcome) correlations.

The processing of data can be defined as the process of collecting raw data and converting it into a format that is useful for subsequent analysis. This phase pertains to the data preparation process that must be completed prior to proceeding with the subsequent phase.

Clean up data by removing null values and unnecessary columns from the dataset. This preparation step involves three main phases: data cleansing, data reduction, and data transformation.

- **Data cleaning**

The preliminary phase of the data processing sequence is data cleansing. This phase includes repairing or removing incorrect, corrupt, malformed, duplicate, or incomplete data from a dataset.

The integration of data from various sources invariably entails the potential for data duplication or mislabeling. If the data is found to be inaccurate, the results and algorithms might not be reliable, even if they seem correct at first.

- **Data reducing**

Data reduction means using less storage space for data. This approach has the potential to enhance storage efficiency, thereby leading to cost reductions. Data reduction includes converting empirical or experimental numerical or alphanumeric information into a rectified, organized, and simplified form.

There are two primary methods for reducing data: (1) the initial step in achieving this objective entails the elimination of invalid data. (2) Secondly, it can be achieved by creating combined data and statistics at different levels of detail for different uses.

- **Data transforming**

Data transformation is the process of changing the format, structure, or value of data, such as importing a database file, an XML document, or an Excel spreadsheet into another file. Transformation is the process of changing the original data source into a formatted version that has been cleaned up and checked.

There are three types of data transformation: constructive (adding, duplicating, and duplicating data), destructive (removing fields and records), aesthetic (standardizing salutations or street names), or structural (renaming, moving, and combining columns in a database).

3.3 Feature Selection

The aim of this step is to improve the predictive accuracy and efficiency of ML models by identifying the most relevant features that influence the forest fires occurrence. This step is performed by GWO, which is a bioinspired algorithm, inspired by the social hierarchy, and hunting behavior of gray wolves.

Here, we present the main steps of the GWO algorithm [12]:

- GWO starts with a randomly initialized population of solutions and iteratively updates the positions of the wolves based on the guidance of the top three individuals (alpha, beta and delta).
- Next, the top three individuals guide the search towards promising regions in the feature space. During the search process, GWO balances exploration and exploitation to avoid local optima and converge to a globally optimal or near optimal feature subset.
- The fitness function typically evaluates both classification performance (e.g. accuracy or F1 score) and feature subset size, often using a weighted sum approach.

For our purpose, GWO serves as a robust metaheuristic optimization algorithm for feature selection tasks. GWO focuses on the most meaningful features, and helps reduce the dimensionality of the dataset, so reducing computational cost and reducing the risk of overfitting.

The following characteristics are retained in most models (RF, LR, and DT) based on the GWO selection process and the nature of the data gathered from [3]:

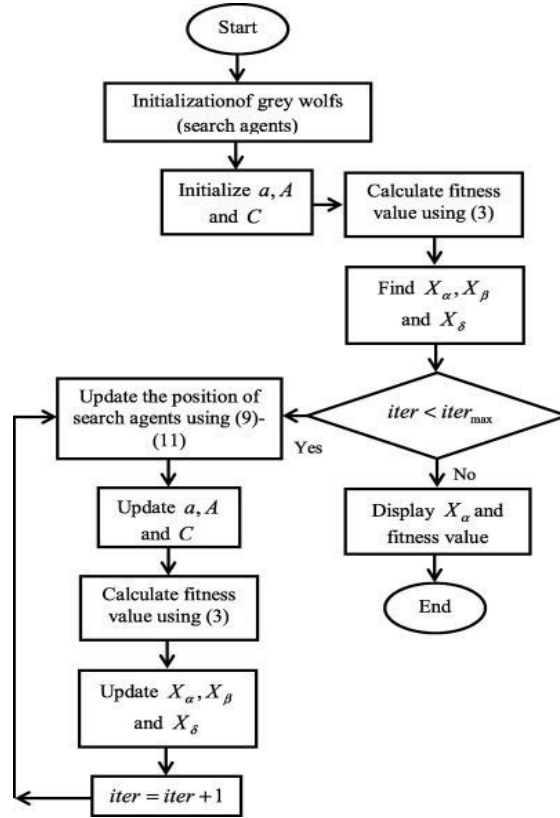


Figure 2: How GWO works [12]

- Temperature
- Relative Humidity (RH)
- Wind Speed (Ws)
- ISI (Initial Spread Index)
- FPMC (Fine Fuel Moisture Code)

3.4 Prediction

This step involves the assessment of the logistic regression model's capacity to predict accurately forest fires.

We note that in this step, we use the three most used and accurate ML models according to the related works such as LR, RF and DT. These models obtained the best accuracy regarding state-of-the-art:

- RF [13] is a ML algorithm that is very popular because of its accuracy, its simplicity and its flexibility. This algorithm can be used for regression and classification problems and is a popular choice for prediction of future behavior and outcomes. It is widely used in many fields, such as banking and e-commerce, to predict future behavior and outcomes.
- LR [14] aims to solve classification problems. Unlike linear regression, which predicts a continuous outcome, it does so by predicting categorical outcomes. In the simplest case, there are two outcomes, which is called a binomial, an example of which is the prediction of whether a tumor is malignant or benign. Other cases have more than two outcomes to classify, in which case it is called multinomial.
- DT [15] is a powerful and intuitive ML algorithm. It is used for both classification and regression tasks. DT is structured like a tree, where each internal node represents a decision based on a particular feature, each branch corresponds to the outcome of the decision, and the leaf nodes represent the final prediction or output (class or

value).

4. Experiment and Evaluation

To validate the effectiveness of our logistic regression-based forest fire forecasting system, we undertook a series of experiments on a set of meteorological data from the regions of Bejaia and Sidi-Bel-Abbes in Algeria.

The dataset was compiled by Abid and Izeboudjen [3] and is available for download from the Kaggle website in CSV format and publicly available on the UCI Machine Learning Repository.

The file size is about 14.3 Kb, comprising 244 instances, 11 meteorological and fire weather index features (Table 1), with a binary target variable: 'fire' (138 instances) and 'not fire' (106 instances). The dataset includes data gathered from two regions in northeast of Algeria, namely Bejaia and Sidi-Bel-Abbes and consists of 11 attributes, plus an output attribute referred to as "class":

Table 1

Dataset attributes

N°	Attribute	value	Description
1	Date	June to September 2012	Date
2	Temp	22 to 42 Celsius degrees	Temperature
3	RH	21 to 90%	Relative Humidity
4	Ws	6 to 29	Wind Speed km/h
5	Rain	0 to 16.8	Outside rain in mm/m ²
6	FFMC	28.6 to 92.5	Fine Fuel Moisture Code
7	DMC	1.1 to 65.9	Duff Moisture Code
8	DC	7 to 220.4	Drought Code
9	ISI	0 to 18.5	Initial Spread Index
10	BUI	1.1 to 68	
11	FWI	0 to 31.1	Fire Weather Index
12	Classes	Target attribute	'Fire' / 'not Fire'

The experiment was performed within a Python environment using Scikit- Learn, Pandas, and Matplotlib libraries. After data preprocessing, we split data into training and test sets. We performed six experiments : RF, RF+GWO, LR, LR+GWO, DT and DT+GWO. We use the metrics : accuracy, Recall and F-score [16] [17].

Table 2

Experiment results

Model	Accuracy	Recall	F-Score
LR	0.90	0.85	0.92
RF	0.75	0.72	0.72
DT	0.83	0.92	0.85
LR + GWO	0.92	0.86	0.93
RF + GWO	0.78	0.73	0.75
DT + GWO	0.85	0.93	0.87

- LR achieved 90% accuracy with 85% recall and 92% F-score. This indicates that the model is highly effective in identifying the occurrence of fires, with a minimum of false positive and false negative results. In addition, the application of GWO has enhanced accuracy from 90% to 92%.
- The DT model achieved 83% accuracy, with 92% recall and 85% F-score. These results demonstrate the model's ability to effectively identify fire events, although it may have a higher false positive rate compared to LR. In addition, the application of GWO has enhanced accuracy from 83% to 85%.
- RF achieved 75% accuracy. The recall is 72%. Although the accuracy is lower than other models, the recall indicates a reasonable ability to detect fire events. In addition, the application of GWO has enhanced accuracy from 75% to 78%.

GWO effectively reduced feature dimensionality. This helped LR and DT to avoid overfitting and allowed RF to focus on the most relevant inputs.

In conclusion, the LR model performs best for the DT model; it offers a good balance between recall and precision. The RF model, although less accurate, can still be useful in certain contexts. We note that the application of GWO for feature selection has improved and optimized each model accordingly.

5. Conclusion

In this paper, we propose a model for the prediction of forest fires. For this purpose, we have used the most three accurate ML algorithm according to the literature. The experiment was completed on meteorological data accessible through the Kaggle platform regarding three ML algorithms: LR, RF and DT.

Experimental results showed that the application of GWO on LR improves the prediction accuracy regarding RF and DT.

In our future work, we plan to use various bioinspired algorithms to improve accuracy.

Declaration on Generative AI

During the preparation of this work, the authors used X-GPT-4 and Grammarly in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

- [1] BARGALI, Himanshu, PANDEY, Aseesh, BHATT, Dinesh, et al. Forest fire management, funding dynamics, and research in the burning frontier: A comprehensive review. *Trees, Forests and People*, 2024, p. 100526.
- [2] BENBAKKAR, Hadj Ali, SOUIDI, Zahira, BARDADI, Abdelkader, et al. Forest fires risk in a semi arid region: a case study in western Algeria. *International Journal of Professional Business Review: Int. J. Prof. Bus. Rev.*, 2024, vol. 9, no 5, p. 1.
- [3] ABID, Faroudja et IZBOUDJEN, Nouma. Predicting forest fire in algeria using data mining techniques: Case study of the decision tree algorithm. In : *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019) Volume 4-Advanced Intelligent Systems for Applied Computing Sciences*. Springer International Publishing, 2020. p. 363-370.
- [4] ELKHRACHY, Ismail, PHAM, Quoc Bao, COSTACHE, Romulus, et al. Sentinel-1 remote sensing data and Hydrologic Engineering Centres River Analysis System two-dimensional integration for flash flood detection and modelling in New Cairo City, Egypt. *Journal of Flood Risk Management*, 2021, vol. 14, no 2, p. e12692.

- [5] PREETI, T., KANAKARADDI, Suvarna, BEELAGI, Aishwarya, et al. Forest fire prediction using machine learning techniques. In : 2021 International Conference on Intelligent Technologies (CONIT). IEEE, 2021. p. 1-6.
- [6] SI, Liqing, SHU, Lifu, WANG, Mingyu, et al. Study on forest fire danger prediction in plateau mountainous forest area. *Natural Hazards Research*, 2022, vol. 2, no 1, p. 25-32.
- [7] SINGH, Kajol R., NEETHU, K. P., MADHUREKAA, K., et al. Parallel SVM model for forest fire prediction. *Soft Computing Letters*, 2021, vol. 3, p. 100014.
- [8] SOUALAH, Louiza, BOUZEKRI, Abdelhafid, et CHENCHOUNI, Haroun. Hoping the best, expecting the worst: Forecasting Forest fire risk in Algeria using fuzzy logic and GIS. *Trees, Forests and People*, 2024, vol. 17, p. 100614.
- [9] SATHISHKUMAR, Veerappampalayam Easwaramoorthy, CHO, Jaehyuk, SUBRAMANIAN, Malliga, et al. Forest fire and smoke detection using deep learning-based learning without forgetting. *Fire ecology*, 2023, vol. 19, no 1, p. 9.
- [10] SHARMA, Sanjeev et KHANAL, Puskar. Forest fire prediction: A spatial machine learning and neural network approach. *Fire*, 2024, vol. 7, no 6, p. 205.
- [11] ZENNIR, Rabah et KHALLEF, Boubaker. Forest fire area detection using Sentinel-2 data: Case of the Beni Salah national forest-Algeria. *Journal of Forest Science*, 2023, vol. 69, no 1, p. 33.
- [12] NEGI, Ganga, KUMAR, Anuj, PANT, Sangeeta, et al. GWO: a review and applications. *International Journal of System Assurance Engineering and Management*, 2021, vol. 12, p. 1-8.
- [13] XU, Chunmei et KONG, Yan. Random forest model in tax risk identification of real estate enterprise income tax. *Plos one*, 2024, vol. 19, no 3, p. e0300928.
- [14] DAS, Abhik. Logistic regression. In : *Encyclopedia of quality of life and well-being research*. Cham : Springer International Publishing, 2024. p. 3985-3986.
- [15] KUSHWAH, Jitendra Singh, KUMAR, Atul, PATEL, Subhash, et al. Comparative study of regressor and classifier with decision tree using modern tools. *Materials Today: Proceedings*, 2022, vol. 56, p. 3571-3576.
- [16] EL BOUHISSI, Houda, TAGZIRT, Dyhia, BOUREDJOUA, Faycal, et al. Health Recommender System for Smart Cities. In : *MoMLeT+ DS*. 2023. p. 334-343.
- [17] BEKKA, Reda, KHERBOUCHE, Samia, et BOUHISSI, Houda El. Distraction detection to predict vehicle crashes: a deep learning approach. *Computación y Sistemas*, 2022, vol. 26, no 1, p. 373-387.