Applied Decision Analytics

Journal homepage: www.ada-journal.org
ISSN: 3104-2945

Unsupervised Phenotyping of Forest Fires Using Clustering of Fire Weather Index and Meteorological Variables

Nataša Milosavljević[1,*] Sretenka Srdić[2]

1 Department of Mathematics and Physics, Institute of Agricultural Engineering, Faculty of Agriculture, University of Belgrade, Belgrade, Serbia
2 Colorado State University, Spur, 4777 National Western Dr., Denver, CO 80216, USA

ARTICLE INFO

Article history:
Received 15 July 2025
Received in revised form 28 September 2025
Accepted 20 October 2025
Available online 4 December 2025

Keywords:
Forest Fire Phenotyping, Wildfire Clustering, Fire Weather Index–based Modeling

ABSTRACT

Due to climate change and human negligence, forest fire problems have become a hot topic in recent years and it is of great importance to be able to prevent them. The aim of this research is to propose a classification model using unsupervised machine learning on the UCI forest fire dataset to identify different fire phenotypes based on the components of the Fire Weather Index (FWI), meteorological variables, weather coding, and burned area characteristics. The model uses K-Means, Gaussian mixture models, DBSCAN (density-based clustering algorithm), and HDBSCAN (Hierarchical Density-Based Spatial Clustering) to optimize the number of clusters. The results obtained show that clustering provides a powerful framework for characterizing forest fire behavior, thereby improving the possibilities for further development in order to suppress and prevent the outbreak of new fires.

1. Introduction

The frequency of wildfires over the past few decades has attracted the attention of modeling researchers to predict and assess the risk of wildfires. Wildfires are driven by complex interactions between meteorological conditions, fuel moisture conditions, seasonal cycles, and long-term drought effects, making their characterization a challenging and multidimensional problem. Traditional assessment systems fail to detect hidden patterns of behavior found in data. With recent advances in data-driven modeling, unsupervised as well as supervised learning have emerged as powerful tools for uncovering hidden patterns in ecological datasets and data.

A significant amount of research has focused on supervised predictions of burned areas, fire spread rate, and ignition probability. Cortez and Morais [1] demonstrated the relevance of meteorological and FWI variables for predicting burned area using machine learning regression techniques on the UCI wildfire dataset. Additional work has explored neural networks, hybrid models, and meteorological and remote sensing data fusion approaches for estimating fire intensity and

* *Corresponding author.*
E-mail address: natasam@agrif.bg.ac.rs

https://orcid.org/0000-0003-4056-089X

behavior [2–4]. While these models provide valuable predictive capabilities, they inherently rely on predefined labels (e.g., measured burned area) and therefore cannot uncover latent structures or intrinsic behavioral regimes within the wildfire phenomenon. Unsupervised approaches offer the potential to uncover natural groupings that arise from environmental drivers rather than assigned targets. Such phenotyping has been shown to be useful in other fields, including climate classification, environmental systems analysis, and ecological modeling. However, there is little research that models unsupervised clustering to characterize wildfires. Existing studies have mostly focused on fixed thresholds or simplified indices, leaving a methodological gap in identifying multivariate fire regimes that are jointly driven by meteorological conditions and FWI components. Addressing the problem of wildfires using unsupervised learning is particularly important because wildfire management requires understanding not only how large a fire can become, but also why and under what environmental conditions different types of behavior occur. Identification of these latent phenotypes supports improved early warning, more precise resource allocation, and improved interpretation of fire-environment interactions. Therefore, the aim of this research is to perform unsupervised phenotyping of wildfires by applying multiple clustering algorithms - including K-Means, Gaussian mixture models, DBSCAN and HDBSCAN - on a feature space composed of meteorological variables, fire weather index components and logarithmically transformed burned areas. The analysis aims to reveal different regimes of wildfire behavior, describe their basic ecological signatures and provide a data-driven basis for operational wildfire risk assessment.

2. Literature Review

Cortez and Morais [1] conducted one of the studies using the UCI Forest Fires dataset and showed that the combination of meteorological variables and FWI components provides significant predictive value for forest fire behavior. On the other hand, Li and Fei [5] extended the research by applying the least squares support vector method to fire prediction, highlighting the importance of nonlinear learning techniques to capture the complex relationships between weather conditions and fire occurrence. These basic studies laid the foundation for more advanced machine learning approaches to forest fire forecasting.

A large number of subsequent studies have investigated different supervised machine learning algorithms for fire detection and prediction. Kumar and Akhlaq [6] developed classification-based methods for fire detection, while Elshewey & Elsonbaty [7] used machine learning techniques to distinguish fires from ambient noise, highlighting the advantages of automated analysis over traditional sensor-threshold approaches. In another contribution, Elshewey [8] compared multiple regression-based machine learning techniques for predicting burned area, demonstrating the sensitivity of performance to model selection and feature construction. Iyer et al. [2] examined several data analysis and machine learning algorithms aimed at predicting random small-scale forest fires, highlighting the importance of detecting such events due to their cumulative ecological impact.

Recent studies include IoT networks, satellite data, big data architectures, and advanced computing frameworks. Gadekar et al. [9] developed a forest fire detection system based on IoT and machine learning that uses real-time sensor anomalies for early warning. Tavakoli et al. [10] addressed the serious problem of class imbalance in fire classification by integrating big data synthesis and class imbalance correction methods, demonstrating improved robustness under skewed data distributions. Xie et al. [11] combined DBSCAN-based non-fire point selection with deep neural networks to improve susceptibility mapping, illustrating the importance of data preprocessing and sample selection in forest fire modeling. In parallel, Ibnat [12] has shown how scalable big data

platforms, such as PySpark, can be used for large-scale environmental forecasting tasks, suggesting potential applicability for operational monitoring of forest fires.

There is a growing body of research comparing the performance of different machine learning models for forest fire forecasting. Kani and Saudia [13] analyzed the accuracy of several machine learning classifiers, emphasizing that model performance is highly dependent on the data structure and target characteristics. Khanjalkar et al. [14] proposed a machine learning-based forecasting model to mitigate the environmental hazards associated with forest fires. Jain et al. [4] presented a new feature selection method for regression and applied it specifically to forest fire forecasting, demonstrating significant improvements in model generalization. Karim et al. [15] applied a midpoint variant of K-means factorial (MKFF) to wireless sensor networks for fire forecasting, demonstrating the utility of clustering to optimize communication and data aggregation in distributed monitoring systems.

Nilashi et al. [16] analyzed type 2 fuzzy systems for biomedical prediction, demonstrating their adaptability to nonlinear noisy environments. Li et al. [17] showed that spiking neural networks can support continuous authentication tasks, highlighting their potential for processing sequential sensor data in real time. Jadhav et al. [3] used satellite data modeling for climate-related predictions, illustrating the cross-domain applicability of remote sensing and machine learning. Zhang et al. [18] developed a generalization-memorization machine for regression, contributing new perspectives on model complexity and performance trade-offs. Raja-Tapiya et al. [19] emphasized the importance of clustering and machine learning for sustainability applications, further reinforcing the relevance of unsupervised approaches in ecological studies.

Most existing research on wildfire analysis has relied on supervised learning, focusing on predicting burned area, classifying fires, or estimating ignition probability [1], [4-11], [13-14]. Although these models have strong predictive capabilities, they inherently depend on labeled outcomes and therefore cannot detect whether natural “types” or phenotypes of fires exist in environmental data. Few studies have used clustering in the context of wildfire analysis, and when clustering is used, it is typically applied to sensor network optimization [15] or pattern filtering [11] rather than to systematically characterize wildfire behavior regimes in a multidimensional feature space. This methodological gap motivates the use of unsupervised learning as a tool to uncover latent wildfire phenotypes, complementing existing supervised approaches and enabling a more nuanced, data-driven understanding of wildfire behavior.

3. Methodology

In order to achieve the best possible separation of data and grouping into clusters that best characterize fires, we used a combination of different methods. Each of them is described in detail in this section (Figure 1.).

3.1 Data Preprocessing

The data used in this work are the publicly available original UCI forest fire dataset https://archive.ics.uci.edu/dataset/162/forest+fires. This dataset contains meteorological variables (temperature, relative humidity, wind speed, precipitation), FWI components (FFMC, DMC, DC, ISI), spatial coordinates (X, Y), time descriptors (month, day), and burned area. Data preprocessing was performed. Categorical time variables (month, day) were transformed into cyclic representations to preserve periodicity. This was achieved by sine-cosine coding, avoiding artificial discontinuities between consecutive time periods (e.g., December and January). The burned area shows a strong

rightward skew, with many zero-area events and a few extremely large fires; therefore, a natural logarithmic transformation was applied to stabilize the variance. All continuous variables were standardized using z-normalization to ensure equal weighting within the distance-based clustering methods. Table 1 summarizes the preprocessing transformations applied to each feature category.

Table 1
Feature categories and preprocessing transformations applied

Feature type	Variables	Transformation	Formula
Meteorological	temp, RH, wind, rain	Standardization	/
Temporal	month, day	Cyclical encoding	If $x \in X$ where X has n elements than $encoding_{\sin} = \sin\left(\frac{2\pi x}{n}\right)$ or $encoding_{cos} = \cos\left(\frac{2\pi x}{n}\right)$
FWI components	FFMC, DMC, DC, ISI	Standardization	$x_{std} = \frac{x - \mu_x}{\sigma_x}$
Fire intensity	Area	Natural logarithm	$area \geq 0, area_{\log} = \ln(area + 1)$

3.2 Feature

The final feature space consisted of 13 standardized variables: four cyclical temporal features (month_sin, month_cos, day_sin, day_cos), four meteorological variables, four FWI components, and log-transformed burned area. This multidimensional representation enabled the clustering algorithms to capture both environmental conditions and underlying fire behavior. The feature engineering process emphasized variables that directly influence ignition potential, fuel moisture, fire spread, and seasonal variation, forming a robust foundation for phenotyping forest fire events.

3.3 Clustering Algorithms

To identify wildfire phenotypes, four unsupervised clustering methods were applied: K-Means, Gaussian Mixture Models (GMM), DBSCAN, and HDBSCAN. K-Means was selected as a primary algorithm due to its simplicity, interpretability, and efficiency in high-dimensional standardized data. GMM was used to capture probabilistic cluster membership, enabling identification of overlapping fire regimes. DBSCAN and HDBSCAN were included to detect noise points and outlier fire events, which often correspond to rare, extreme-intensity fires. These methods collectively allow the discovery of compact, probabilistic, and density-based phenotypes. All figures associated with clustering analysis follow CDF Letters quality standards, ensuring clear readability and appropriate label sizing, as demonstrated in Figure 2.

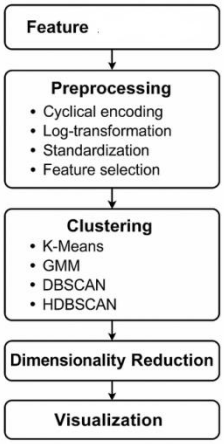

Fig. 1. Workflow diagram of the implemented preprocessing and clustering methodology

3.4 Cluster Validation

For better clustering, we used the silhouette score measure to determine how well separated the data points are within the clusters.

$SilhouetteScore = \frac{x_i - y_i}{max(x_i, y_i)}$, where x_i is the average distance from the data point to all other points in the same cluster and y_i is the average distance from the data point to all points in the nearest neighboring cluster.

A high score indicates that the data points are close to their own cluster and far from others, while a negative score suggests that the point may be in the wrong cluster. It is used to assess the quality of the clustering results and can help determine the optimal number of clusters for an algorithm like K-Means. Silhouette scores were computed for *k* values ranging from 2 to 8, and the configuration with the highest score was selected for final clustering. Density-based algorithms do not require preset values of *k*, and instead evaluate intrinsic data density to identify cluster cores and noise points. Silhouette results for tested values of *k* are summarized in Table 2.

Table 2.
Silhouette coefficient results for different values of k ranging from 2 to 8.

K	Silhouette score
2	0.332019
3	0.336792
4	0.14188
5	0.142131
6	0.145045
7	0.143331
8	0.150131

3.5 PCA-Based Visualization

Principal Component Analysis (PCA) was applied to the standardized dataset to obtain two-dimensional Figure 2 and three-dimensional Figure 3 projections of the clusters. PCA allowed clear visualization of separation between identified wildfire phenotypes. Figures 2 and 3 representing PCA

plots were generated in high resolution and saved with appropriately sized labels to satisfy journal formatting expectations.

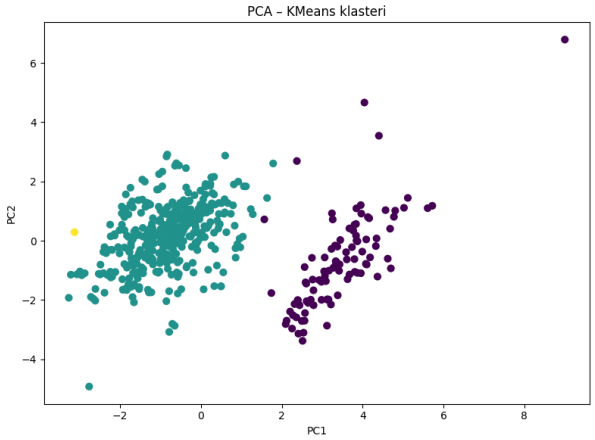

Fig. 2. Two-dimensional PCA visualization of K-Means clustering results

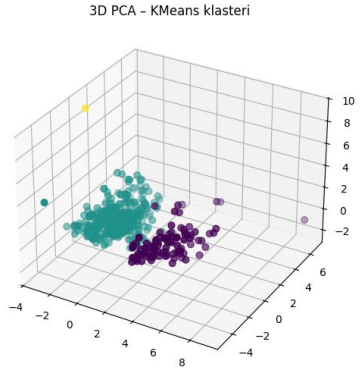

Fig. 3. Three-dimensional projections of the clusters

3.6 Cluster Validation Metrics

The optimal number of clusters (*k*) for centroid-based algorithms was determined using silhouette analysis. The silhouette coefficient provides a quantitative measure of internal cohesion and separation between clusters. The silhouette score as a function of *k* was evaluated for a range of values from 2 to 8, and the configuration with the maximum coefficient was selected for subsequent analysis.

All clustering results, cluster labels, centroid coordinates, silhouette metrics, and summary statistics were exported to an Excel file containing multiple structured sheets. In addition, high-resolution figures were automatically saved in PNG format to ensure compatibility with CDF Letters' publication requirements. To adhere to journal standards, all figures were generated with Calibri font, readable annotations, and minimal unused frame area. Figures that appeared side-by-side were used only when their combined width comfortably fit within the page, as illustrated in Figure 4 and Figure 5.

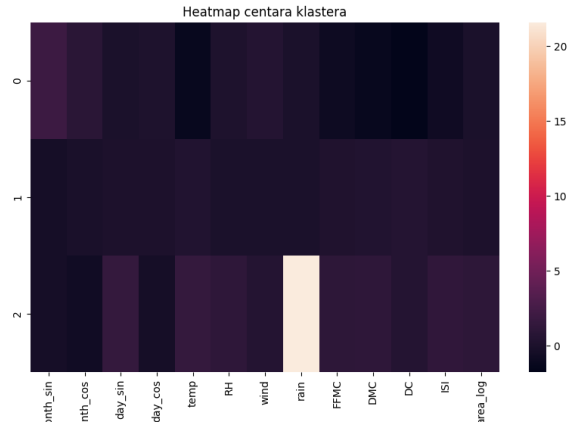

Fig. 4. Cluster centroid heatmap illustrating feature contributions

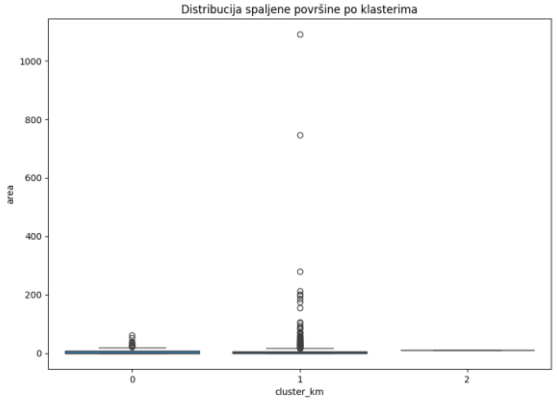

Fig. 5. Boxplot distribution of burned area across wildfire phenotypes

4. Results

From the results obtained, clustering analysis revealed a clear structure within the multidimensional feature space, indicating that wildfire behavior in the Montesinho region can be grouped into distinct environmental phenotypes. Silhouette estimation identified k = 3 as the optimal number of clusters for centroid-based methods, providing the greatest balance between intra-cluster cohesion and inter-cluster separation. The PCA representation in Figure 2 shows clearly identifiable cluster boundaries in the reduced two-dimensional space, confirming that the standardized set of features carries significant discriminative information related to wildfire dynamics.

K-Means clustering produced three compact and well-separated wildfire phenotypes. Cluster 0 represents fires associated with high temperatures, low relative humidity, and elevated fine fuel moisture deficit (high FFMC), accompanied by moderate to high values of burned area. These conditions characterize high-fire risk environments characterized by increased ignition potential and rapid rate of spread. Cluster 1, in contrast, corresponds to low-intensity micro-fires, where higher humidity, lower FFMC, and limited fuel dryness result in minimal burned area. Cluster 2 reflects the moderate-to-high intensity fire phenotype, characterized by extremely high DC values, indicating long-term seasonal drought conditions even when short-term moisture indicators are less extreme. The burned area distributions shown in Figure 4 further highlight the contrast in behavior between phenotypes, with clusters showing distinctly different profiles of mean and maximum burned areas.

The GMM yielded similar structural divisions, albeit with more relaxed probability boundaries between clusters. The overlap observed in the GMM membership suggests that certain fire events exist at the transition between clusters, influenced by gradual changes in meteorological conditions rather than clear thresholds. This behavior is typical of ecological systems with smooth environmental gradients. The consistency between K-Means and GMM confirms that the identified phenotypes are robust to algorithmic assumptions.

Density-based clustering algorithms provided additional insights. DBSCAN identified a small number of points as noise or outliers, and these events correspond to rare fires of extreme intensity characterized by unusually high burned areas and severe drought conditions. These cases, which typically represent less than 5% of all observations, highlight the importance of incorporating density-based techniques when characterizing wildfire behavior, as centroid-based methods tend to absorb outliers into nearby clusters. HDBSCAN further refined this separation, producing stable cluster cores and a larger forest region, demonstrating its utility for isolating anomalous fire behavior. A heatmap visualization of the K-Means centroids provides a consolidated view of the feature-level contributions to each phenotype. Cluster 0 displays uniformly elevated FFMC, DMC, and temperature values, consistent with surface fuel dryness. Cluster 1 exhibits the lowest ISI and FFMC, indicating a potential for slow spread and lower ignition probability. Cluster 2 is dominated by a strong DC signal, emphasizing the role of long-term seasonal drought rather than short-term atmospheric conditions. This differentiation suggests that different fire regimes arise from a combination of short-term and long-term climate drivers.

Spatial analysis (Figure 6) revealed that the geographic distribution of clusters was relatively uniform across the study area, suggesting that topography played a minimal role in differentiating fire phenotypes within this dataset. Instead, variability appears to be predominantly driven by meteorological and fuel-related conditions. This observation is consistent with the design of the dataset, which was collected from a relatively homogeneous region in terms of vegetation type and elevation.

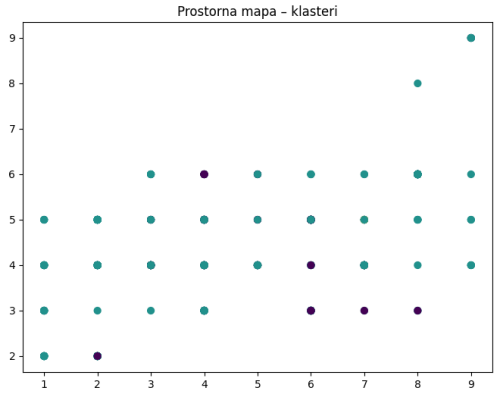

Fig. 6. Spatial analysis

Overall, the results indicate that unsupervised clustering effectively uncovers three dominant wildfire phenotypes in the Montesinho region:

i. High-intensity fires driven by elevated fine-fuel dryness and low humidity,
ii. Low-intensity micro-fires associated with moist conditions and minimal fuel availability, and
iii. Moderate to large fires are influenced primarily by long-term deep-layer drought.

These phenotypes provide a data-driven foundation for understanding how environmental conditions shape wildfire behavior and offer practical value for early warning systems, operational risk classification, and strategic resource planning.

5. Conclusions

The present study demonstrates that unsupervised learning provides an effective framework for characterizing distinct wildfire behavioral regimes based on meteorological variables, Fire Weather Index (FWI) components, and log-transformed burned-area measurements. Using a combination of K-Means, Gaussian Mixture Models (GMM), DBSCAN, and HDBSCAN, three dominant wildfire phenotypes were identified within the UCI Forest Fires dataset. These phenotypes correspond to high-intensity fires driven by elevated fuel dryness and low humidity, low-intensity micro-fires occurring under moist atmospheric conditions, and moderate to large fires influenced predominantly by long-term seasonal drought.

The consistency between centroid-based and probabilistic clustering approaches highlights the robustness of these phenotypes. Density-based methods further isolated rare extreme-intensity events, emphasizing their importance as outliers in wildfire behavior analysis. Visual inspection through PCA projections and burned-area distributions confirmed clear separation among the phenotypes and demonstrated the relevance of both short-term and long-term drought indicators.

The findings illustrate that latent wildfire regimes can be extracted directly from multivariate environmental data without relying on supervised labels. This contributes a novel perspective to wildfire characterization and supports the development of more adaptive and data-driven fire risk assessment frameworks. Future work may extend this analysis using larger multisite datasets, integrate remote-sensing indicators, or incorporate temporal sequence modeling to better capture dynamic fire evolution.

Acknowledgement

This research was supported by Ministry of Science, Technological development, and Innovations, the grant agreement registration for the Faculty of Agriculture, No. 451-03-137/2025-03/200116.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Cortez, P., & Morais, A. D. J. R. (2007). A data mining approach to predict forest fires using meteorological data. *Proceedings of the 13th Portuguese Conference on Artificial Intelligence*, Guimarães, Portugal.

[2] Iyer, V., Iyengar, S. S., Paramesh, N., Murthy, G. R., & Srinivas, M. B. (2011). Machine learning and data mining algorithms for predicting accidental small forest fires. In Proceedings of SENSORCOMM 2011: The Fifth International Conference on Sensor Technologies and Applications (str. 116–121).

[3] Jadhav, M. M., Agarwal, P., Umadevi, B., Khatibi, A., Akhila, N., Sandeep, K. S., & Banerjee, S. (2024). Climate change prediction in sustainable healthcare systems for biodiverse ecosystem based on satellite data modelling. Remote Sensing in Earth Systems Sciences, 7(4), 283–293. https://doi.org/10.1007/s41976-024-00120-4

[4] Jain, N., Ghosh, S., & Ghosh, A. (2024, December). A novel feature selection method for regression and its application for prediction of forest fires. In *2024 IEEE India Geoscience and Remote Sensing Symposium (InGARSS)* (pp. 1–4). IEEE.

[5] Li, E., & Fei, Y. (2016). Prediction of forest fires based on least squares support vector machine. Hans Journal of Data Mining, 6(1), 15–27. https://doi.org/10.12677/hjdm.2016.61003

[6] Kumar, A., & Akhlaq, A. (2022). Forest fire detection. International Journal of Health Sciences, 6(S2), 7504–7510. https://doi.org/10.53730/ijhs.v6nS2.6807

[7] Elshewey, A. M., & Elsonbaty, A. A. (2020). Forest fires detection using machine learning techniques. *Journal of Xi'an University of Architecture & Technology*, 12(9). ISSN No : 1006-7930

[8] Elshewey, A. M., & Elshewey, A. M. (2021). Machine learning regression techniques to predict burned area of forest fires. *International Journal of Soft Computing*, 16(1), 1–8.

[9] Ghadekar, P., Amune, A., Sayyed, M., Patil, S., Vidhale, A., & Zanwar, N. (2024). Forest fire detection using IoT and machine learning. In *Artificial Intelligence and Information Technologies* (pp. 374–380). CRC Press. https://doi.org/10.1201/9781032700502-60

[10] Tavakoli, F., Naik, K., Zaman, M., Purcell, R., Sampalli, S., Mutakabbir, A., Lung, C., & Ravichandran, T. (2024, February). Big Data Synthesis and Class Imbalance Rectification for Enhanced Forest Fire Classification Modeling. In 16th International Conference on Agents and Artificial Intelligence (ICAART 2024, Volume 2) (pp. 264–275). SciTePress. https://doi.org/10.5220/0012363000003636

[11] Xie, L., Zhang, R., Lv, J., Shama, A., & Yang, Y. (2025). Enhancing forest fire susceptibility mapping in Xichang City, China using DBSCAN-based non-fire point selection integrated with deep neural network. *Geomatics, Natural Hazards and Risk*, 16(1), 2443465. https://doi.org/10.1080/19475705.2024.2443465

[12] Ibnath, M. H. (n.d.). *Forest Cover Type Prediction Using PySpark and Machine Learning Models: A Big Data Approach*.

[13] Kani, D. C. J., & Saudia, S. (2023, January). Analysis on the performance of machine learning models for forest fire prediction. In *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1–5). IEEE.

[14] Kanjalkar, J., Kanjalkar, P., Golegaonkar, P., Bissa, D., Borade, I., Ratnalikar, P., & Bora, T. (2024, July). Machine learning-based forest fire forecasting: Mitigating environmental hazards. In 2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE) (pp. 1–5). IEEE. https://doi.org/10.1109/ICSPCRE62303.2024.10675176

[15] Karim, R., Zahedi, M., De, D., & Das, A. (2024). MKFF: Mid-point K-means based clustering in wireless sensor network for forest fire prediction. Microsystem Technologies, 30(4), 469–480. https://doi.org/10.1007/s00542-023-05578-8

[16] Nilashi, M., Abumalloh, R. A., Ahmadi, H., Samad, S., Alyami, S., Alghamdi, A., Alrizq, M., & Yusuf, S. Y. M. (2024). Accuracy analysis of Type-2 fuzzy system in predicting Parkinson's disease using biomedical voice measures. International Journal of Fuzzy Systems, 26(4), 1261–1284. https://doi.org/10.1007/s40815-023-01665-0

[17] Li, Y., Sun, X., Yang, Z., & Huang, H. (2024). SNNAuth: Sensor-based continuous authentication on smartphones using spiking neural networks. IEEE Internet of Things Journal, 11(9), 15957–15968. https://doi.org/10.1109/JIOT.2024.3349533

[18] Zhang, S., Bai, L., & Wang, Z. (2025, April). Generalization memorization machine for regression. In *International Conference on Computer Application and Information Security (ICCAIS 2024)* (Vol. 13562, pp. 45–56). SPIE. https://doi.org/10.1117/12.3061042

[19] Raya-Tapia, A. Y., López-Flores, F. J., Ramírez-Márquez, C., & Ponce-Ortega, J. M. (2025). Machine learning and clustering for a sustainable future: Applications in engineering and environmental science. Studies in Computational Intelligence, 1233. Springer. https://doi.org/10.1007/978-3-032-03876-0