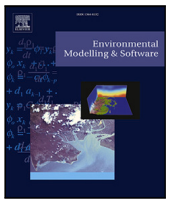
Environmental Modelling and Software 193 (2025) 106578

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft



AutoML-Fire: Automated machine-learning approach to predict forest fires

Saurabh Toraskar [a], Adil Khan [a], M. Niranjannaik [a], Abhilash Singh [a,b,*], Kumar Gaurav [a,**]

[a] Fluvial Geomorphology and Remote Sensing Laboratory, Department of Earth and Environmental Sciences, Indian Institute of Science Education and Research Bhopal, Bypass road, Bhopal, 462066, Madhya Pradesh, India

[b] School of Mathematics, Faculty of Engineering and Physical Sciences, University of Leeds, Leeds, LS2 9JT, United Kingdom

ARTICLE INFO

Keywords:
Forest fire
Automated machine learning
Bayesian optimisation
SHAP
SMOGN

ABSTRACT

Forest fires pose a serious threat to the environment. Their frequency and intensity have increased in recent decades due to climate change and heightened anthropogenic interference. About 21.7% of India's total land is covered by forests, which play a crucial role in biodiversity, ecology, and the livelihoods dependent on these ecosystems. However, 36% of these forested regions are prone to frequent and devastating fire. Given this vulnerability, predicting forest fires is essential for minimising damage. In this first pan-India study, we predict the forest fire occurrence in the most vulnerable regions across India using a dataset spanning from 2003 to 2018, incorporating variables such as cloud cover, elevation, forest cover fraction, humidity, NDVI, population, soil moisture, temperature, wind speed, and precipitation. We partitioned the data into four clusters based on spatial proximity to capture regional patterns. SHAP (SHapley Additive exPlanations) values were utilised to enhance model interpretability and provide insights into socio-technical complexities unique to different regions. We analysed the Partial Dependence Plots (PDP) to capture the trend of forest fires with individual features. The challenge of data imbalance, often encountered in natural hazard prediction, was addressed using the Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) algorithm, which balances regression data. Selecting appropriate machine-learning models and adeptly tuning their hyperparameters is a complex process that requires domain expertise. To address this, we proposed an automated machine-learning (AutoML) framework that utilises Bayesian optimisation to return a best-performing, finely-tuned model. The "AutoML-FIRE" model exhibited robust performance, with R values between 0.73 and 0.85 and RMSE values ranging from 3.40 to 6.09, outperforming all considered benchmarking algorithms. Furthermore, uncertainty analysis and spatial distribution analysis were conducted to validate the model's stability. Our analysis demonstrates that the AutoML-FIRE model is robust, exhibits broad applicability for national-scale fire risk assessment, and enables notifications to authorities and local communities regarding impending fire events.

1. Introduction

Forest fires pose a severe threat to the environment and human life (Robinne and Secretariat, 2021). Their occurrences drastically reduce the productivity of land ecosystems, soil degradation, and release of significant amounts of carbon into the atmosphere (Sannigrahi et al., 2020; Pérez-Cabello et al., 2012; Seibert et al., 2010; Venkatesh et al., 2020). The rise in global temperatures has been identified as a key driver of the increasing incidence of forest fires. This trend has been starkly illustrated by recent catastrophic fire events in regions such as California, Australia, and the Amazon, that destroyed several square kilometres of forest area (Attri et al., 2020; Boer et al., 2020). Globally, about 80% instances of forest fires constitute in savannahs and grasslands across North and South America, Australia, Africa, and South Asia (Schultz et al., 2008).

In recent years, India has experienced a significant uptick in the occurrences of forest fires. In many instances, they are primarily driven by human activities, including land preparation for agriculture, deforestation, controlled burns, and the collection of forest products (Reddy et al., 2019). In South Asia, about 32.2% of forest fire instances are from India alone. Their occurrences are concentrated in specific hotspot regions (Ahmad and Goparaju, 2019). The regional variability

* Corresponding author at: School of Mathematics, Faculty of Engineering and Physical Sciences, University of Leeds, Leeds, LS2 9JT, United Kingdom.
** Corresponding author.
E-mail addresses: toraskar21@iiserb.ac.in (S. Toraskar), adil23@iiserb.ac.in (A. Khan), niranjannaik@iiserb.ac.in (M. Niranjannaik), a.singh4@leeds.ac.uk, abhilash.iiserb@gmail.com (A. Singh), kgaurav@iiserb.ac.in (K. Gaurav).

https://doi.org/10.1016/j.envsoft.2025.106578
Received 30 November 2024; Received in revised form 14 April 2025; Accepted 15 June 2025
Available online 10 July 2025

of forest fires in India is mainly due to geographical diversity and varied topography, climate conditions, vegetation, and forest types. This poses a challenge for developing a universal predictive model for India to predict forest fires. Identifying the key variables influencing forest fires and determining their relative importance is another complex task. It is further complicated by the imbalanced nature of forest fire data, where events are rare but have severe consequences.

Machine learning (ML) models can be a very useful tool to predict forest fires accurately. However, the success of these models largely depends on the quality of the input data and careful selection and tuning of the hyperparameters of the model. Further, addressing the data imbalance and regional variability is challenging. To overcome these limitations,this study proposes a generalised automated machine-learning model (“AutoML-FIRE”) for pan-India by leveraging Bayesian optimisation to predict forest fires. This study also seeks to investigate the underlying dynamics of causative factors of forest fires at both national and regional scales. This is achieved through SHapley Additive exPlanations (SHAP) analysis, which provides a detailed interpretation of feature importance and contributes to understanding the socio-technical dimensions specific to different regions. This study enhances predictive capabilities for forest fires while providing a robust framework to address regional variability and data imbalance across India.

2. Related works

Machine learning models to predict forest fires have gained considerable momentum, ranging from fire spread prediction to the deployment of early warning systems (Wijayanto et al., 2017; Radke et al., 2019). A noteworthy case study by Yang et al. (2021) introduced a machine learning model named Agni, capable of predicting forest fires up to one month in advance. This model achieved an area under the receiver operator characteristic (ROC) curve greater than 0.81. However, the model did not incorporate key meteorological parameters like precipitation and soil moisture, nor did it account for human activities, which could have enhanced its robustness.

To improve the accuracy and efficiency of fire management strategies, Xu et al. (2022) developed a model for predicting forest fire spread using a cellular automata (CA) framework combined with least squares support vector machines (LSSVM). This LSSVM-CA model accounted for the effects of adjacent wind on fire spread, but it heavily relied on vegetation data, potentially limiting its accuracy in regions with sparse or inaccurate vegetation information.

Vega-Garcia et al. (1996) utilised neural networks to predict human-induced wildfire occurrences. Logistic regression analysis served as the “domain expert” to identify important input variables, resulting in a model that correctly predicted 85% of no-fire observations and 78% of fire observations. However, manually adjusting hyperparameters such as learning rate and network architecture is time-consuming and requires significant expertise. An automated framework could potentially improve the accuracy and reliability of such models by optimising hyperparameters automatically.

Cortez and Morais (2007a,b) explored the development of a cost-effective, real-time forest fire prediction tool using readily available meteorological data. They employed support vector machines (SVM) and random forests (RF), alongside four distinct feature selection setups. Although this model was tested on real-world data from the northeast region of Portugal, its ability to predict large fires could be enhanced by incorporating additional information such as vegetation type and firefighting interventions.

The influence of demographic factors on forest fire predictions has also been studied. Kang et al. (2020) made progress in fire risk prediction by integrating road and population data, leading to more granular spatiotemporal patterns. Similarly, Kim et al. (2019) demonstrated a direct relationship between population density and forest fire risk through the incorporation of socioeconomic trends in machine learning models.

Despite significant global advancements, research on forest fires prediction in India remains limited. Babu et al. (2023) demonstrated the effectiveness of ensemble-based machine learning models for fire prediction in the Western Ghats region, showcasing the potential of advanced techniques in specific ecological zones. Similarly, Saha et al. (2023) focused on the Ayodhya hills, identifying forest fire susceptibility zones using RF, multivariate adaptive regression splines (MARS), and deep learning neural networks (DLNN), contributing valuable insights into regional fire risk. However, these studies are confined to particular regions, and there is a noticeable absence of comprehensive research on a pan-India level. Such studies are crucial for understanding the broader patterns and drivers of forest fires across diverse ecological and climatic zones in India. A pan-India approach is essential for stakeholders to develop robust, scalable forest management strategies and policy frameworks that can address the varying fire risks across the country’s extensive and diverse landscapes.

Overall, the main challenge in forest fire prediction systems appears to be integrating diverse indicators to produce accurate, consistent, and computationally efficient predictions. Manual selection of algorithms can introduce biases, and optimising hyperparameters across different models often lacks a standardised approach. AutoML frameworks provide a solution by automating these processes, reducing biases, and enhancing model performance (Singh et al., 2022, 2024; Kumar et al., 2024). They have demonstrated promising performance in forest fire prediction tasks, but these applications have been focused on the classification domain (Zhang and Pan, 2024; Su et al., 2024; Kong, 2024). In this study, we propose an AutoML framework for forest fire count prediction, representing the first such comprehensive study conducted on a pan-India scale. The implementation of AutoML enhances hyperparameter tuning, effectively addressing the gaps identified in earlier studies. In addition, we tackle the issue of imbalanced data distribution using the SMOGN algorithm and broaden the scope of previous research by incorporating both socio-economic factors and meteorological parameters.

3. Study area

The Indian landmass has been classified into 16 distinct climate zones (Beck et al., 2018). The northern and northeastern areas, with temperate climates and monsoonal rains, sustain dense vegetation, increasing fire risk in the dry periods. Conversely, arid and semi-arid zones in the western and central regions are particularly fire-prone due to hot, dry conditions that intensify fuel combustibility. The varying topography from the Himalayas and Western Ghats to the Deccan Plateau, creates a heterogeneous landscape where fire dynamics vary considerably. For example, steep slopes in the northern and western regions can expedite fire spread, while flat plains influence ignition patterns and fire intensity. Vegetation types, such as tropical dry deciduous, evergreen, montane, and thorn forests, add to this complexity, as each exhibits unique fire susceptibility due to distinct fuel loads and seasonal dryness (Roy and Purohit, 2018).

According to a report of the Indian State of Forest Report (ISFR) 2021, forests occupy 21.7% of the geographical area of which the states of Madhya Pradesh, Arunachal Pradesh, Chhattisgarh, Odisha, and Maharashtra have significant forested areas that can be susceptible to forest fires. To conduct an in-depth analysis of forest fire risks, this study employs a grid-based approach, segmenting India into $0.25° \times 0.25°$ cells. This enables a granular examination of fire-prone regions, supporting more targeted prediction and mitigation efforts across diverse Indian landscapes.

4. Material and methods

4.1. Data

We compiled the forest fire occurrences in India from 2003 to 2018. We have downloaded the forest fire data from the Forest Survey of India (FSI) website (https://fsiforestfire.gov.in/index.php). This dataset includes forest fire data points derived from the Fire Information for Resource Management System (FIRMS), which utilises MODIS (Moderate Resolution Imaging Spectroradiometer) observations. Corresponding to the forest fire events, we obtained the daily minimum and maximum air temperature (Tas_{min} [°C] and Tas_{max} [°C]) measured at a height of 2 m above the Earth's surface at spatial resolution 1° × 1° from the Climate Data Store (Copernicus Climate Change Service, Climate Data Store, 2021). We have downloaded the daily gridded precipitation [mm/day] at a spatial resolution of 1°×1° from the Global Precipitation Climatology Project (GPCP dailyv1.3) and CDS (Adler et al., 2020). We have also used the agrometeorological indicators such as daily mean wind speed [m/s] at 10 m above the surface, cloud cover fraction over 24 h, and relative humidity (%) at 09:00 AM local time at 2 m above the surface. These indicators are obtained from the ERA5 reanalysis product at a high spatial resolution of 0.1° × 0.1° (Boogaard et al., 2020). We downloaded the 30 arc seconds global digital elevation model (DEM) GTOPO30 from the U.S. Geological Survey's (USGS) Center for Earth Resources Observation and Science (EROS) (Earth Resources Observation and Science Center, U.S. Geological Survey, U.S. Department of the Interior, 1997) (see Table 1). We acquired the Normalised Difference Vegetation Index (NDVI) at a daily time scale with a spatial resolution of 0.05° × 0.05° from the National Oceanic and Atmospheric Administration (NOAA) Climate Data Record (CDR) of AVHRR Surface Reflectance (Vermot, 2022). We obtained the annual forest cover fraction (FCF) data [%], at a spatial resolution 0.5° × 0.5°, from ICDC, CEN, University of Hamburg (DiMiceli et al., 2022). Population density is also an important parameter that influences fire risk. To incorporate this in our study, we obtained the population count data from the Gridded Population of the World, Version 4 (GPWv4) dataset by CIESIN at Columbia University (Center for International Earth Science Information Network-CIESIN-Columbia University, 2018). This provides the population counts for 2000, 2005, 2010, 2015, and 2020 at a spatial resolution of 0.25° × 0.25°, consistent with national censuses and population registers.

4.2. Data processing

The fire count data is a point measurement, each point represents the location of a single fire occurrence. We convert this to the grid data format at a spatial resolution of 0.25° × 0.25° by summing the number of fire points within each grid cell for each day. Accordingly, using nearest neighbour interpolation, we have resampled all the input variables to a uniform resolution at 0.25° × 0.25°. The elevation data (GTOPO30) was originally available at a 1 km spatial resolution, we have converted it to a 0.25°×0.25° grid resolution, by averaging all the values lying inside the grid of targeted resolution, providing a constant value for each grid cell. For the variables having a frequency of record less than daily, we convert them to daily time resolution by assigning the same value for each day of the corresponding period. Gholami et al. (2021) have shown that historical rainfall data can be used to predict fire occurrences. We have used the average rainfall of previous year to capture the long-term effects of rainfall on the occurrence of forest fires.

Forest fire occurrence is governed by a complex interplay of meteorological, topographic, and ecological factors that exhibit strong spatial dependencies. Many forest fires are caused by human activity, with such anthropogenic sources being particularly localised, especially in community-driven countries like India. This trend further reinforces the concept of spatial clustering. Clustering grids based on spatial proximity provides a physically meaningful way to segment the study area into distinct fire-prone regions with shared environmental conditions that drive fire behaviour. To perform clustering, we first identified the grids having the top 10%, 20%, and 30% of the total fire count in the forest fire time series data from 2003–2018. After initial computations, it was found that the top 30% grids provided the most promising results and were subsequently used for further analysis. We segment this fire count data into 4 clusters (Fig. 1b) based on their spatial proximity using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. We set the clustering parameters $\epsilon = 1.3$ and $\min_{samples} = 10$. The ϵ is the maximum distance between two samples used as a criterion for a cluster. The $\min_{samples}$ specifies the minimum number of samples required to form a cluster.

4.3. Feature importance and association

SHAP (Shapley Additive exPlanations) introduced by Lundberg and Lee (2017) quantify the effect of a feature i on the model's prediction by training the model twice. First including the feature i ($f_{S\cup\{i\}}$) and then without the feature i (f_S). The term S represents a subset of the feature set T, excluding feature i ($S \subseteq T \setminus \{i\}$). The difference between these two model outputs, $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$, measures the impact of feature i on the model's prediction. SHAP values are computed as weighted averages of these differences (Eq. (1)).

$$\phi_i = \sum_{S\subseteq T\setminus\{i\}} \frac{|S|!\,(|T|-|S|-1)!}{|T|!}\left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right] \tag{1}$$

SHAP is an additive feature attribution method, meaning that the sum of the effects of all feature attributions approximates the original model output, as shown in Eq. (2).

$$f(x) = \mathbb{E}[f(x)] + \sum_{i=1}^{T} \phi_i \tag{2}$$

In this study, we used Tree SHAP implementation from the *SHAP* Python library to compute Shapley values (Lundberg et al., 2018). The SHAP feature importance method identifies the most influential features based on the magnitude of their absolute Shapley values (Molnar, 2022). To obtain a global understanding of feature importance, we calculate the average absolute Shapley value for each feature from Eq. (3).

$$I_i = \frac{1}{N_{obs}} \sum_{j=1}^{N_{obs}} |\phi_i^{(j)}| \tag{3}$$

where N_{obs} denotes the number of observations.

Additionally, it is crucial to ensure that input features exhibit minimal correlation for optimal machine learning model performance. Highly correlated features can destabilise the model, leading to inaccurate estimations (Singh et al., 2021, 2024). To analyse feature correlations, we have plotted the feature association matrix, which quantifies the similarity between decision rules in partitioning data, measuring how closely potential splits align with the final optimal split during tree growth (Bhadani et al., 2024). This is done to ensure that multicollinearity does not adversely affect the model's performance.

4.4. Feature sensitivity

We employed Partial Dependence Plots (PDPs), as proposed by Friedman (2000), to visualise the relationship between the target variable and individual features. PDPs offer insight into how the predicted target value changes as a function of a single feature while averaging out the effects of all other features. This approach allowed us to observe how the feature impacts the model's predictions on average. The mathematical formulation for the partial dependence of the ith feature can be written according to Eq. (4).

$$\hat{f}(x_i) = \frac{1}{N_{obs}} \sum_{j=1}^{N_{obs}} f(x_i, x_{T\setminus\{i\}}^{(j)}) \tag{4}$$

Table 1
Detailed description of datasets used in this study, such as data source, and spatio-temporal resolution.

Data	Units	Source	Temporal resolution	Spatial resolution
Fire count (Target variable)	Count	Forest Survey of India (FSI)	Time of occurrence	Point data
Wind speed Cloud cover Humidity	m/s Dimensionless %	ERA 5 Reanalysis product Agrometeorological indicators (Sourced from CDS)	Daily	0.1° × 0.1°
Minimum temperature Maximum temperature	°C	BERKEARTH dataset (Sourced from CDS)	Daily	1° × 1°
Rain	mm	GPCP daily v1.3 (Sourced from CDS)	Daily	1° × 1°
Soil moisture	kg/m^3	NASA GES DISC	Daily	0.25° × 0.25°
NDVI (Normalised Difference Vegetation Index)	Dimensionless	NOAA CDR	Daily	0.05° × 0.05°
Forest cover fraction	%	Dataset created by ICDC, CEN University of Hamburg	Annual	0.5° × 0.5°
Population	Count	Gridded Population of the World (GPWv4)	Constant value	0.25° × 0.25°
Elevation	m	GTOPO	Constant value	0.0083° × 0.0083°

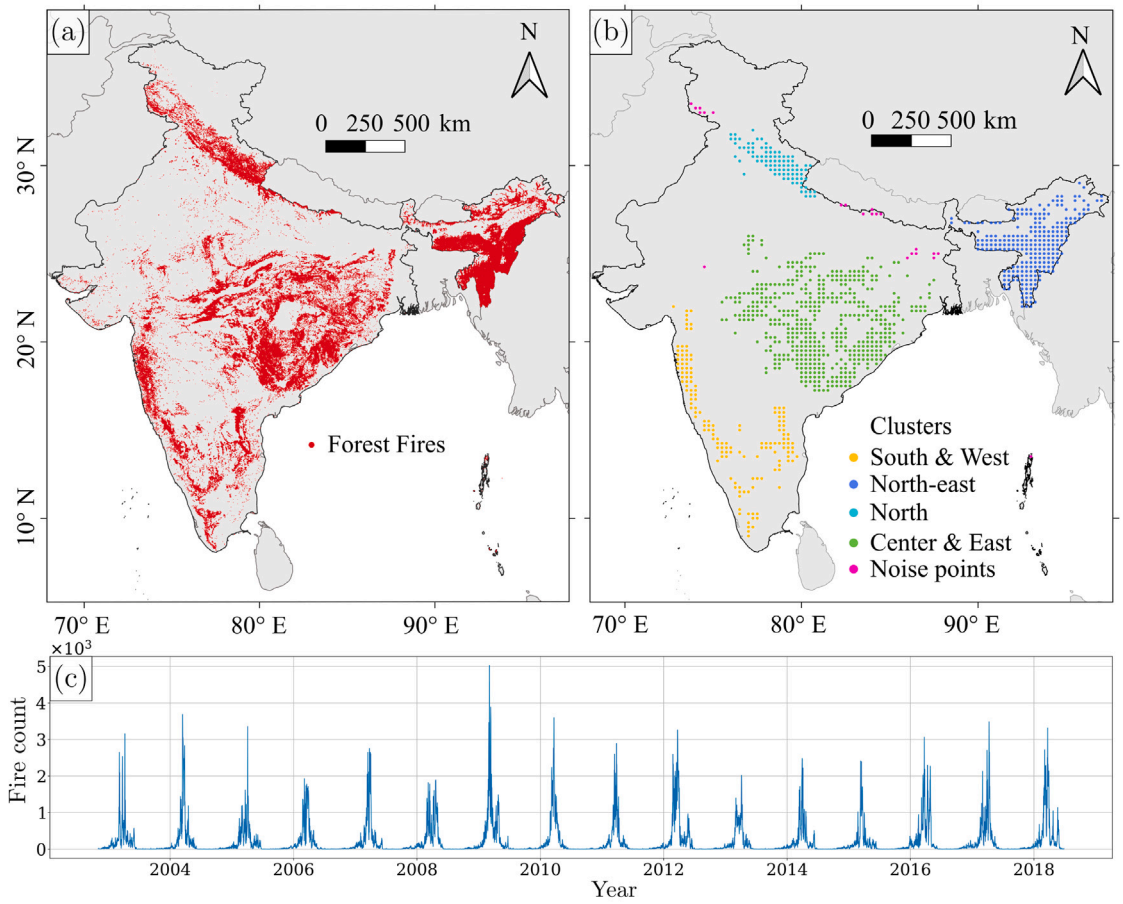


Fig. 1. The study area map shows the occurrence and distribution of forest fire events in India. (a) shows the spatial distribution of forest fires from 2003 to 2018. (b) shows the centre point of the 0.25° × 0.25° grid with the top 30% of forest fire events. Further, these grids are classified as four clusters, namely South & West (yellow), Center & East (green), North-East (blue), and North (cyan) regions, and noise grids from the clusters are shown with purple points. (c) shows the temporal variability of daily forest fire counts from 2003 to 2018.

where $x^{(j)}_{T\setminus\{i\}}$ represents the fixed values of all other features for the jth observation, and N_{obs} is the total number of observations in the dataset.

In addition to PDPs, we utilised Individual Conditional Expectation (ICE) plots, introduced by Alex Goldstein and Pitkin (2015), to further investigate feature sensitivity at the individual observation level. While PDPs provide an averaged effect across all observations, ICE plots disaggregate this effect, showing how predictions change for each observation as a particular feature varies. This allows for the examination

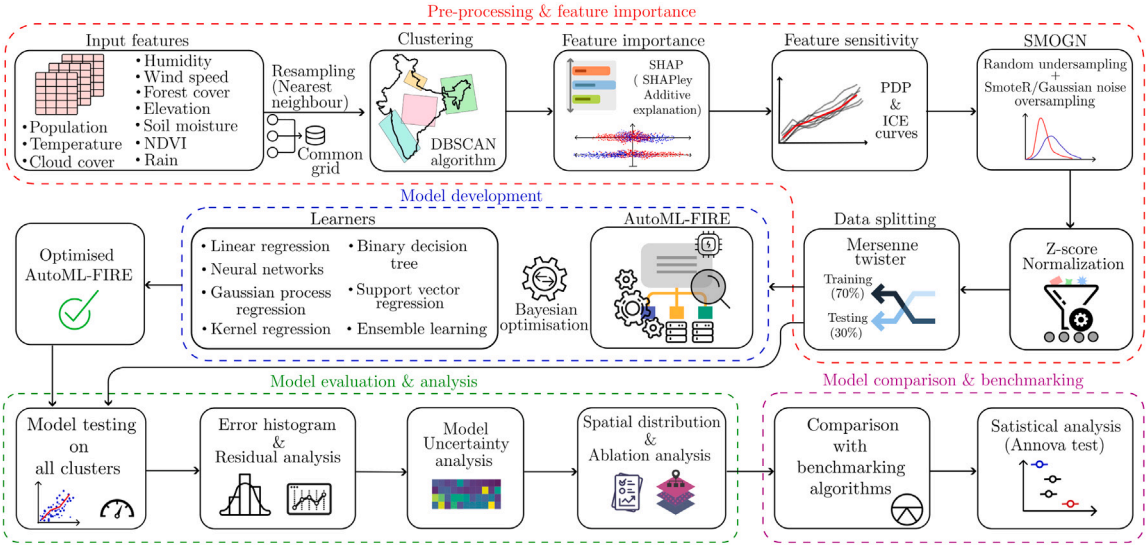


Fig. 2. The flowchart illustrates the detailed methodology used for predicting forest fire counts. The processes inside the red dashed box represent the preparation of features from input datasets. The blue dashed box depicts the model development. The following two boxes at the bottom represent model evaluation and comparison with benchmark algorithms.

of heterogeneous effects, revealing patterns that may be hidden in PDPs. The ICE plot for the jth observation is computed as follows:

$$\hat{f}_j(x_i) = f(x_i, x^{(j)}_{T\setminus\{i\}}) \tag{5}$$

ICE plots help to highlight how individual predictions vary in response to changes in a specific feature, whereas PDPs offer a global view by displaying the average effect of the feature across the entire dataset.

4.5. Data generalisation

Fire count shows high frequency of lower values, leading to a skewed distribution. Details on this are explained in the Fig. 1 of the supplementary material. To address this, we used the Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) proposed by Branco et al. (2017). It combines random under-sampling (Torgo et al., 2013, 2015) with two oversampling techniques; SmoteR (Torgo et al., 2013) and the introduction of Gaussian noise (Branco et al., 2016). Initially, a relevance function, as proposed by Torgo and Ribeiro (2007) and Ribeiro (2011), is defined using the target variable sample distribution to map the target variable values to a relevance scale ranging from 0 to 1. The SMOGN algorithm categorises the data into two using the relevance function; BinsR and BinsN. The SMOGN partitions in BinsN undergo random under-sampling, whereas the rare samples in BinsR are subject to over-sampling. For each seed example in BinsR, synthetic cases are generated using either SmoteR or Gaussian noise, depending on the distance to the selected k-nearest neighbour. If the neighbour is within a "safe" distance, SmoteR is employed; otherwise, a Gaussian noise is introduced. The safety threshold is set at half the median distance between the seed and other examples within the same partition. Finally, we normalised all eleven features and utilised them in training for the prediction of forest fire count using the AutoML-Fire algorithm. To do so, we have used a Python implementation of SMOGN provided by Kunz (2020). The k parameter was set 5, where k specifies the number of neighbours to consider for interpolation used in over-sampling, and the *samp_method* (sampling method) was set to 'balance' to ensure the synthetic data points do not deviate much from real-world data points.

4.6. AutoML-FIRE

Model selection is a critical phase in predicting any target variable. This process not only requires identifying an appropriate model but also involves the systematic evaluation of a wide range of hyperparameters, as these hyperparameters significantly affect the model's performance. Given the large number of possible hyperparameter configurations, manual tuning becomes both time-consuming and impractical. Furthermore, this manual approach lacks a standardised method for ensuring optimal model selection. Fine-tuning hyperparameters to maximise performance for a specific target variable also demands a deep understanding of the model architecture and the nature of the target variable, often making the task highly challenging and unfeasible.

To streamline the process of model selection and hyperparameter tuning, we utilise a standardised Automated Machine Learning (AutoML) framework that leverages Bayesian optimisation to identify and fine-tune the most effective model for the given data. The algorithm driving this framework is outlined in Algorithm 1. The next subsection details the workings of the Bayesian optimisation technique used within our framework, while the following subsections describe the individual models evaluated in this study. A detailed methodology of AutoML-FIRE is shown in Fig. 2.

4.6.1. Bayesian optimisation

Bayesian optimisation (BO) employs Bayes' theorem to identify the maximum or minimum of an objective function efficiently. In contrast to conventional methods like random search or grid search, BO accelerates the optimisation process by incorporating knowledge from previous evaluations. This approach constructs a probabilistic model, known as a surrogate function, based on past observations to approximate the true objective function. A commonly used surrogate model in BO is the Gaussian Process (GP), as expressed in Eq. (6).

$$f(x) \sim \mathcal{GP}(\mu(x), C_v(x, x')) \tag{6}$$

Here, $f(x)$ follows a Gaussian distribution with a mean function $\mu(x)$ and covariance $C_v(x, x')$. The covariance between any two points, x and x', is modelled using a radial basis function (RBF) kernel according to Eq. (7):

$$C_v(x, x') = \exp\left(-\frac{1}{2}\|x - x'\|^2\right) \tag{7}$$

Algorithm 1 Pseudocode for AutoML-FIRE algorithm.

```
Inputs: fire_dataset (Input dataset), fire_target (Response variable)
Outputs: optimal_model (Best predictive model), optimal_hyperparams (Best hyperparameters)
function AutoML-FIRE(fire_dataset, fire_target)
    optimal_model ← None
    highest_score ← −∞
    train_data, validation_data ← split_data(fire_dataset, split_ratio = 0.7)
    model_set ← [FFNN, SVR, GPR, RF, Boosting, BDT, LR, KR]                ▷ Model pool
    num_threads ← number of processors available for parallel computing
    Parallel Execution                                                     ▷ Parallelising model training
    for each model_type ∈ model_set in parallel using num_threads do
        optimal_hyperparams ← Bayesian_Tuning(model_type, train_data, fire_target)
        score_total ← 0
        for fold ∈ cross_validation_folds(train_data, num_folds) do
            trained_model ← model_type.train(fold.train_data, fire_target, optimal_hyperparams)
            fold_score ← model_type.evaluate(fold.val_data, fire_target)
            score_total += fold_score
        end for
        average_score ← score_total/num_folds
        if average_score > highest_score then
            highest_score ← average_score
            optimal_model ← trained_model
        end if
    end for
    return optimal_model
end function
function Bayesian_Tuning(model_type, train_data, fire_target)
    hyperparam_space ← model_type.get_hyperparameter_space()
    function Objective(hyperparams)
        candidate_model ← model_type.train(train_data, fire_target, hyperparams)
        val_score ← model_type.evaluate(validation_data, fire_target)
        return val_score
    end function
    optimal_hyperparams ← Bayesian_Optimisation(Objective, hyperparam_space)
    return optimal_hyperparams
end function
function Bayesian_Optimisation(Objective, hyperparam_space)
    optimal_hyperparams ← optimise(Objective, hyperparam_space)
    return optimal_hyperparams
end function
```

This surrogate function is significantly easier to optimise compared to the original objective function, enabling BO to perform more efficient hyperparameter tuning. The next set of hyperparameters is selected by an acquisition function, which identifies the point in the search space where the surrogate function predicts the most promising performance. This approach ensures that BO iteratively improves the model's performance by prioritising hyperparameters with the highest expected improvement.

4.6.2. Feed-forward neural network

Feed-forward neural networks (FFNN) operate without loops, allowing information to flow in a single direction (Singh and Gaurav, 2024). The fundamental architecture of an FFNN consists of an input layer, an output layer, and hidden layers, each containing a specific number of nodes serving as data-processing units. Each layer is governed according to Eq. (8)

$$y_j = f(\sum_{i=1}^{i} w_{ij} \cdot x_i + b_j) \tag{8}$$

where y_j is the output of the jth node of the current layer, x_i is incoming values from the ith node of the previous layer, w_{ij} are the weights assigned to the connections between the nodes of the current layer and previous layer, b_j is the bias of the current layer and $f()$ is the activation function such as Sigmoid, ReLU, tanh, etc.

4.6.3. Gaussian process regression

Gaussian process regression proposed by Rasmussen et al. (2004) is a non-parametric model, which uses a Bayesian regression approach. It takes into account the uncertainties of the model's output. For ease of understanding, only one input feature is assumed for explanation. Before considering any data, GPR defines a prior distribution with the assumption that the function values $f = [f(x_1), f(x_2), \ldots, f(x_n)]$ have a joint multivariate normal distribution given the training set $D = \{x_i, y_i\}_{i=1}^{n_{obs}}$, where y_i is adjusted so that the mean is 0.

$$f \sim \mathcal{N}(0, K) \tag{9}$$

where $K = K_{X,X} = \kappa(x_i, x_j|\tau)$ is the covariance matrix also called the kernel function, and τ is the set of its hyper-parameters. The most common kernel is the Radial Basis function (RBF) kernel given according to Eq. (10)

$$\kappa(x_i, x_j|\tau) = \sigma^2 exp(-\frac{1}{2}(\frac{\|x_i - x_j\|}{l_s})^2) \tag{10}$$

where σ is the noise term, and l_s is the length scale. GPR then updates the prior distribution, assuming the target values are noisy observations of a true function f:

$$y_i = f(X_i) + \epsilon_i \tag{11}$$

where ϵ_i is the noise term and is assumed to follow a normal distribution, $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. For making predictions, we assume that the training

targets y and the function that predicts the values for test data f^* follows a multivariate normal distribution

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{X,X} + \sigma_\epsilon^2 I & K_{X,X^*} \\ K_{X^*,X} & K_{X^*,X^*} \end{bmatrix} \right) \tag{12}$$

From the above joint distribution, the posterior distribution of the test function values f^*, given the training data D, is a normal distribution with mean μ^* and covariance Σ^*.

$$f^*|(X^*, D) \sim \mathcal{N}(\mu^*, \Sigma^*) \tag{13}$$

where

$$\mu^* = K_{X^*,X} \left[K_{X,X} + \sigma_\epsilon^2 I\right]^{-1} \mathbf{y} \tag{14}$$

$$\Sigma^* = K_{X^*,X_*} - K_{X^*,X} \left[K_{X,X} + \sigma_\epsilon^2 I\right]^{-1} K_{X,X^*} \tag{15}$$

where μ^* represents the prediction of the function for test inputs and Σ^* represents the uncertainty of these predictions.

4.6.4. Support vector regression

Support vector regression (SVR) modifies the well-known SVM for regression tasks where the aim is to find a classifying hyperplane (Vapnik et al., 1996). For non-linear relationships, where the data is not linearly separable in the original feature space, kernels are employed to map the data into a higher dimensional space. SVR is then applied to this transformed data in the higher dimension feature space. The hyperplane separating the data can be obtained from Eq. (16)

$$\vec{\omega} \cdot \vec{x} - b = 0 \tag{16}$$

where $\vec{x}$ is the position vector of the hyperplane, $\vec{\omega}$ is the weight vector normal to the plane, which has the dimensions of that of $\vec{x}$, and b is the intercept or bias term.

Given the training set with m features $\{\vec{x}_i, y_i\}_{i=1}^{n_{obs}}$ where $\vec{x}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]$ we estimate $\vec{\omega}$ where $\vec{\omega} = [\omega_1, \omega_2, \ldots, \omega_m]$ according to Eq. (17)

$$\text{Minimise:} \quad \frac{1}{2}\|\vec{\omega}\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{17}$$

subject to:

$$\begin{aligned} y_i - \vec{\omega} \cdot \vec{x}_i - b &\leq \epsilon + \xi_i \\ \vec{\omega} \cdot \vec{x}_i + b - y_i &\leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned} \tag{18}$$

where ϵ is the margin of tolerance which is the range where no penalty is given to error, ξ_i and ξ_i^* is called the slack variables which allow some error beyond ϵ but puts a penalty on the misclassification, C is called the box constraint which is a regularisation parameter (prevents overfitting) which puts weights on this misclassification penalty. This value determines the trade-off between the ϵ and the amount up to which deviations larger than ϵ are tolerated.

4.6.5. Ensemble learning

Ensemble learning models combine the predictions of multiple weak learners, such as decision trees. Doing so enables the ensemble learning models to produce better results than individual models and also reduces the chances of overfitting, which is a frequent challenge encountered in decision trees. The different methods of creating assemblages are bagging, boosting, and stacking.

In boosting, the model is sequentially trained to correct the predecessor's errors. The method used in this study is Least-Squares Boosting (LSBoost) (Breiman, 2001; Hastie et al., 2009). The algorithm initialises by training some weak learners on the data, or often it just takes the mean of all observations $\bar{y}_i$. Then it calculates the errors between the observed and the initialised values $\bar{y}_i$. The error is calculated using the mean squared error (MSE) method. Then, it iteratively trains and fits a learner, such as a decision tree, in which the target variable is the error of the ensemble of all the previous learners to minimise the MSE. Thus, the prediction is given by

$$\hat{y}_i = \bar{y}_i + \eta f(\vec{x}_i) \tag{19}$$

where $f(\vec{x}_i)$ is the aggregated response from all the weak learners and η is the learning rate.

4.6.6. Kernel regression

Similar to SVR, kernel regression maps the original feature space to a higher dimension using the kernel trick (Hainmueller and Hazlett, 2014). Unlike SVR, the kernel regression fits a linear model to this transformed data in higher dimension space by using the least square regression method. Thus, by using a kernel, the algorithm can find a linear model in a transformed, high-dimension space, which is equivalent to a nonlinear model in the original, lower-dimension space.

4.6.7. Binary decision tree

A decision tree is a hierarchical model that recursively partitions the input space into two by making binary decisions at each node based on feature values (Breiman, 2017). The goal is to find an optimal split for a node t according to some splitting criteria (Loh and Shih, 1997). The splitting criterion utilised here is a mean squared error (MSE), other well-known criteria are Gini impurity, Variance reduction, etc.

Given a training set with m features $\{X_i, y_i\}_{i=1}^{n_{obs}}$ where $X_i = [x_1, x_2, \ldots, x_m]$, for finding the optimal split at node t, for each feature x_j we calculate the weighted MSE and probability that an observation is in node t by

$$\epsilon_T = \sum_{i \in T} \omega_i (y_i + \bar{y}_T)^2 \tag{20}$$

$$P(T) = \sum_{i \in T} \omega_i \tag{21}$$

where T is the set of observations in the node t, ω_i is the weight of ith observation, $w_i = \frac{1}{n}$ if not specified, y_i is the target value for ith observation and $\bar{y}_T$ is the mean of all observations in node t.

Before splitting, all the values in each feature x_j are sorted in ascending order. Subsequently, observations in node t are split into the left child node (t_L) and right child node (t_R) according to a splitting candidate. BDT evaluates all potential splits for each predictor x_i and selects the split that maximises the reduction in MSE (ΔI)

$$\Delta I = P(T)\epsilon_T - P(T_L)\epsilon_{t_L} - P(T_R)\epsilon_{t_R} \tag{22}$$

where $P(T_L)$ and $P(T_R)$ are the probabilities of observations being in the left and right child nodes, respectively, and ϵ_{t_L} and ϵ_{t_R} are the weighted MSEs of the left and right child nodes, respectively.

4.6.8. Linear regression

Linear regression aims to model a linear relationship between the independent variable $\vec{x}$ and the dependent variable $\vec{y}$ using statistics (Hastie et al., 2009). If we are given a training set with m features $\{\vec{x}_i, y_i\}_{i=1}^{n_{obs}}$, where $\vec{x}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]$ then we have to find the best fit linear equation given by

$$f(\vec{X}) = \vec{y} = \vec{\beta} \cdot \vec{X} + \epsilon \tag{23}$$

where $\vec{X} = [1, x_1, x_2, \ldots, x_m]$ and $\vec{\beta} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_m]$ is a vector of estimated coefficients and ϵ is the error or noise term. The objective is to minimise the sum of square residuals (SSR):

$$SSR = \sum_{i=1}^{n_{obs}} y_i - \vec{\beta} \cdot \vec{X}_i \tag{24}$$

where $\vec{X}_i = [1, x_{i1}, x_{i2}, \ldots, x_{im}]$ and $\vec{\beta}$ is estimated by Ordinary Least Squares (OLS) by using Eq. (25):

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{25}$$

where $X = [\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_{n_{obs}}]$, $\hat{\beta} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_m]^T$, and $Y = [y_1, y_2, \ldots, y_{n_{obs}}]^T$. Predictions are done by substituting the new vector $\vec{X}_0$ in Eq. (23).

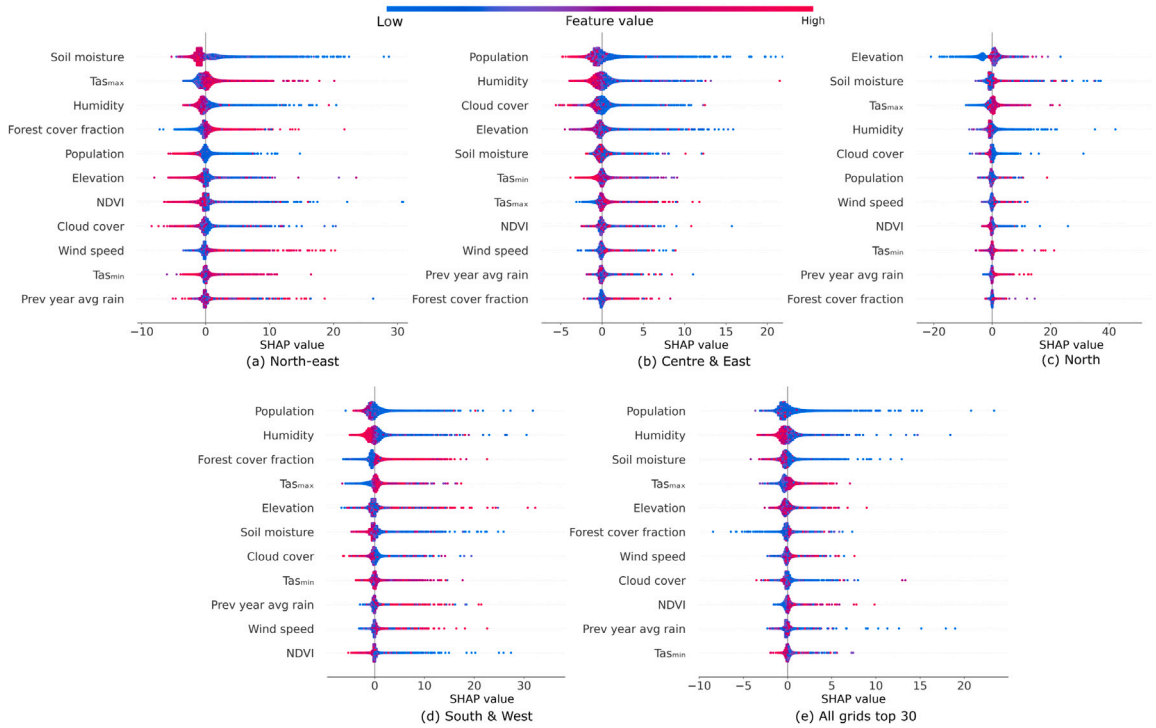


Fig. 3. SHAP summary (Bee swarm) plot showing all features for each cluster's top 30% of grids individually. In this plot, the features are arranged in decreasing order of importance in each subplot, highlighting the key drivers of forest fire predictions across regions.

5. Results

5.1. Feature importance and association

Fig. 3 presents the SHAP (SHapley Additive exPlanations) bee swarm plots for all clusters, illustrating the feature contributions across different regions. In these plots, each point represents an individual observation, with the x-axis showing the magnitude of a feature's contribution to the prediction, while the colour gradient indicates the scaled feature values. For understanding, blue-shaded points on the positive side of the x-axis indicate that lower feature values contributed to an increase in the predicted forest fire count, whereas red-shaded points on the positive side suggest that higher feature values played a role in increasing the predicted fire count. Features are arranged in descending order of their importance within each subplot, enabling easy comparison.

Across the Centre & East, South & West clusters, and the combined top 30% of grids, population emerges as the most influential feature in predicting forest fires. As shown in Fig. 3a, b, and d, higher population values correspond to a decrease in forest fire occurrences. This inverse relationship likely reflects the fact that high population densities are generally concentrated in urban areas where forests are scarce. However, this trend is not observed in the North cluster (Fig. 3c), where the population does not exhibit a clear association with fire counts. This exception may be due to the distinct socioeconomic dynamics of the North region, where forests are preserved even in areas with relatively high population density. The local economy's dependence on eco-tourism and the influx of tourists into forested areas might obscure the direct impact of permanent population data on fire counts in this region. Authorities should implement regulations to control the environmental impact of eco-tourism, particularly in fire-prone forest areas, and raise awareness about fire risks among tourists to mitigate the potential for human-caused fires in these vulnerable regions.

Another key feature contributing to fire prediction across all clusters is humidity. Ranked second in importance in the Centre & East, South & West, and combined grids (Fig. 3b, d, e), third in the North-East (Fig. 3a), and fourth in the North (Fig. 3c), humidity plays a critical role in forest fire dynamics. Notably, soil moisture is the top-ranking feature in the North-East cluster (Fig. 3a), underscoring its importance in fire prediction for that region. Lower humidity values, along with soil moisture and maximum temperature, show a strong positive contribution to fire occurrence. This aligns with the physical conditions required for fire ignition, where low moisture and high temperatures create a dry environment conducive to wildfires (Jain et al., 2022; Mina et al., 2023; Chaparro et al., 2015). Since humidity, soil moisture and temperature are crucial factors in fire dynamics, fire management policies should incorporate real-time monitoring systems for these variables to better predict fire risk.

Elevation presents a more mixed impact across clusters, particularly in the North-East and North regions. In the North-East, lower elevation values are linked to higher fire frequencies, likely due to the prevalence of traditional slash-and-burn agricultural practices among the indigenous population residing in these lower-altitude areas (Dhar et al., 2023). In this region, policies should focus on promoting and incentivising sustainable land-use alternatives alongside education programs on the environmental and fire risks of slash-and-burn, which can reduce its prevalence and mitigate fire risks in the region. Conversely, in the North cluster, the relationship between elevation and fire frequency is more complex. While very low elevation values tend to correlate with fewer fires, high elevation areas do not exhibit a consistent trend. This may be due to the presence of lowland plains with sparse forest cover within this cluster, which skews the overall pattern, leading to fewer fires at lower elevations without a clear relationship at higher altitudes.

To further quantify feature importance, we computed the average absolute SHAP values for each feature. Fig. 4 shows the ranked importance of all features across different clusters, highlighting the variability in feature importance based on regional characteristics. This visualisation offers a comprehensive view of how feature relevance shifts between clusters, providing insights into the regional drivers of forest fire occurrence.

Finally, to ensure model robustness, we evaluated feature interdependencies by constructing a feature association matrix (Fig. 5).

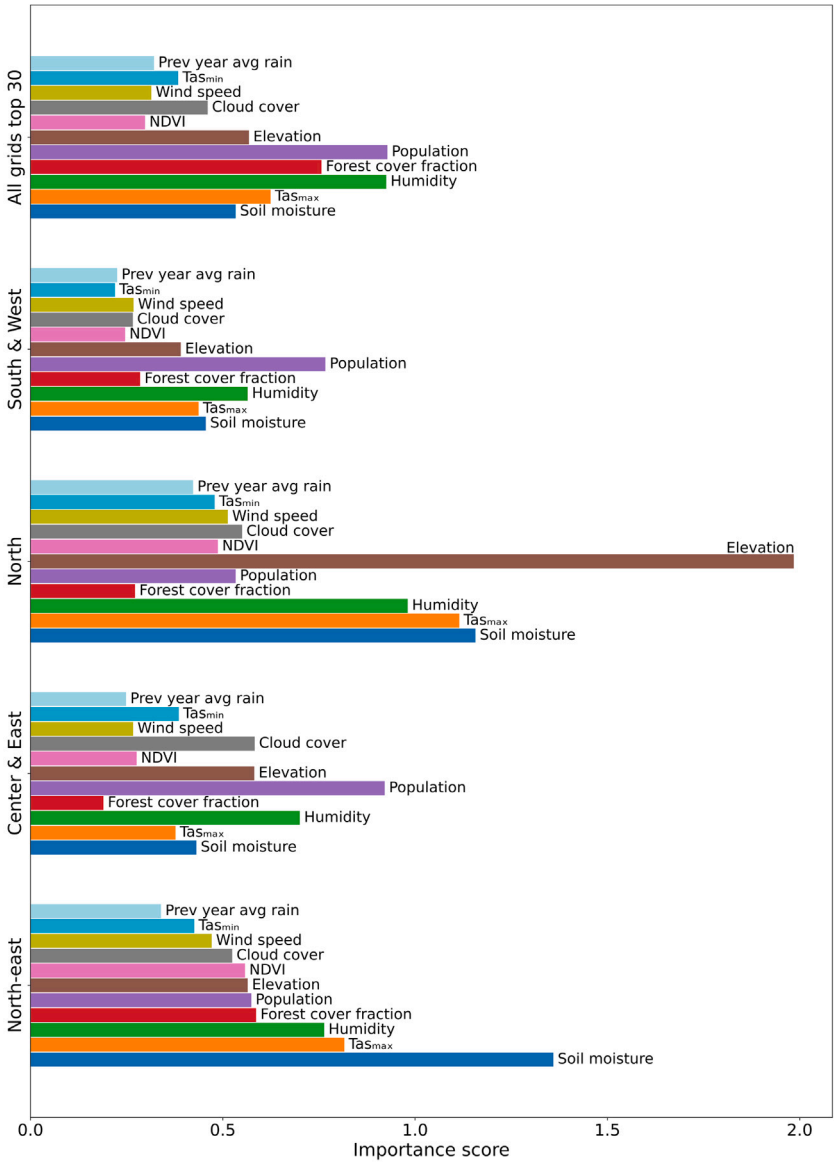


Fig. 4. The feature importance graph for each cluster shows various feature's importance scores, calculated by averaging the absolute SHAP values from the bee swarm plot. This plot depicts the factors that drive the forest fire count.

Machine learning models perform best when input features are independent, as multicollinearity between features can degrade model accuracy. The feature association matrix allows us to identify any significant correlations between features, which, if present, could diminish the predictive performance of the model. This step ensures that the input features used for prediction are not heavily correlated, preserving the integrity of the model's predictions.

5.2. Feature sensitivity

To understand how individual features impact forest fire counts, we plot the PDP and ICE curves for the top 30% grids across India. These visualisations reveal the overall trends in forest fire behaviour as key features vary.

The cloud cover trend in Fig. 6a shows that high cloud cover corresponds to a reduced number of forest fires. This is expected, as increased cloud cover diminishes the intensity of solar radiation reaching the Earth's surface, thereby reducing the likelihood of fire ignition (Pfister et al., 2003). For elevation, the pattern is more complex, displaying an undulating trend for most of the range but showing a sudden spike at around 4000 m, followed by a steep increase. This spike is likely due to the presence of natural forests in the Himalayan grids, which are known for their high forest density (Fig. 6b).

The relationship between forest cover fraction and fire count, as depicted in Fig. 6c, shows a clear upward trend. Greater forest cover provides more available fuel, thereby increasing the potential for forest fires. Similarly, the NDVI exhibits an inverse relationship with forest fire counts–lower NDVI values, which indicate deteriorating plant conditions and lower fuel moisture content (FMC), lead to higher fire activity (Chuvieco et al., 2004; Maselli et al., 2003) (Fig. 6e). However, the increase in fire count after an NDVI value of 0.5 likely reflects the higher biomass (fuel) availability in densely vegetated regions, combined with seasonal or drought-induced drying, which increases fire susceptibility.

Both humidity and maximum temperature are significant drivers of forest fires. As shown in Fig. 6d and i, decreasing humidity and increasing temperatures are closely associated with a rise in fire counts, a trend that aligns with existing research (Jain et al., 2022; Mina et al., 2023).

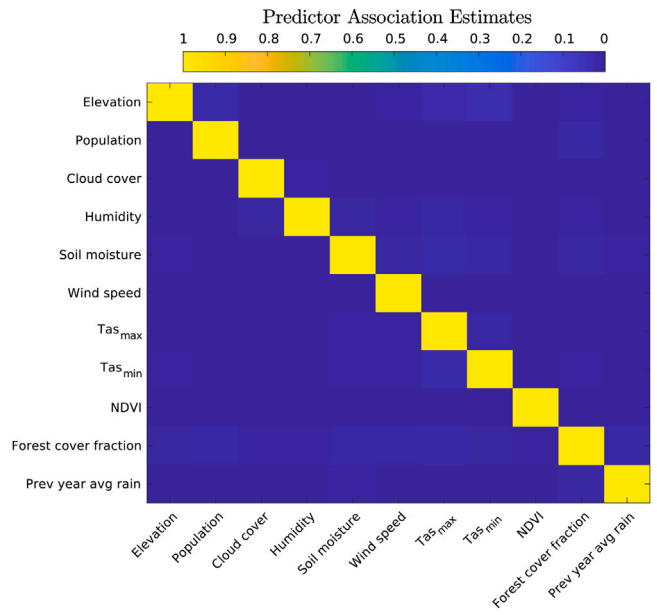


Fig. 5. The feature association estimates heat map illustrates the correlation of all the features used for fire count prediction.

The relationship between population and forest fires, shown in Fig. 6f, appears counterintuitive. While the PDP line suggests an opposite trend, it is important to note that high population values (represented by the blue dots) are primarily associated with urban areas, which generally lack forests and, therefore, have fewer fires.

In Fig. 6g, the trend between the previous year's average rainfall and fire count is positive, indicating that higher rainfall promotes tree growth, which can subsequently provide more fuel for fires (Toledo et al., 2011).

The role of litterfall, which refers to dead plant material such as leaves, twigs, and bark, is also critical. This material is highly flammable and can accelerate the ignition and spread of fires. As seen in Fig. 6i and j, litterfall is positively correlated with maximum temperature and negatively correlated with minimum temperature (Wang et al., 2021), which explains the observed trends. The negative trend in Fig. 6j further underscores how hotter and drier conditions are conducive to forest fires.

Finally, soil moisture and wind speed exhibit notable relationships with fire occurrence. Lower soil moisture is strongly linked to increased fire counts, as drier conditions favour fire ignition and spread (Fig. 6h) (Chaparro et al., 2015). Wind speed shows a decreasing trend up to around 2 m/s, followed by a steady increase (Fig. 6k). This is in line with previous studies, which indicate that fire front speeds are faster in wind conditions ranging from 2 to 6 m/s, and that in mountainous regions, weak winds combined with local topography can heavily influence fire behaviour (Beer, 1991; Brotak, 1991).

5.3. Performance of the AutoML-FIRE

The AutoML-FIRE framework was developed independently for all clusters using their respective datasets. To assess the performance of AutoML-FIRE, we used 70% of the data for training the model and the remaining 30% for testing purposes. The number of data points in the training and testing sets for all the clusters is given in Table 2. To select the optimal hyperparameters, Bayesian optimisation was employed to iteratively enhance the model's performance on the training data. This process continues until the maximum number of iterations is reached, which ensures that convergence is achieved. For our model, the maximum number of iterations was automatically set to 250, based on criteria such as convergence checks to balance exploration and exploitation, optimising the objective function effectively. During this process, the algorithm refines the hyperparameter selection, ensuring convergence of the loss function within the allotted iterations. On completion, the best-performing hyperparameters and the corresponding model are returned. The optimisation results are shown in Fig. 7.

Table 2
Number of data points present in the training and testing split of all clusters.

Cluster	Total	Training	Testing
North-east	48 385	33 870	14 515
Center & East	50 276	35 193	15 083
North	10 527	7369	3158
South & West	16 487	11 541	4946
All grids top 30	129 949	90 964	38 985

The best-performing model for each cluster, returned by AutoML-FIRE, was evaluated on the testing data to assess its robustness and generalisation capabilities. The evaluation metrics, correlation coefficient (R), root mean square error (RMSE), and bias were calculated across the clusters. R values ranged from 0.69 to 0.85, demonstrating a strong correlation between predicted and observed fire counts. This indicates that the model effectively captures fire-prone conditions. High correlation values suggest that the model can differentiate between periods of high and low fire activity, enabling proactive decision-making. RMSE values range from 3.40 to 6.09, representing the average deviation between predicted and actual fire counts per grid. In practical terms, this means that for the given grid, the predicted fire counts typically deviate from the observed values by approximately 3 to 6 fires per day. This level of error is reasonable given the large area of a grid of $0.25° \times 0.25°$ resolution and the inherent variability of fire occurrence, which is influenced by stochastic factors such as sudden ignition events (e.g., lightning, human activity) that are not always captured in meteorological and environmental data. Bias is calculated by taking the average difference between the predicted and observed values and ranges from 0.9 to 1.46, which indicates only a slight overprediction of less than 1.5 forest fire count per grid per day, which is acceptable given the range of the target variable. The evaluation metrics for each cluster, along with the number of data points in the testing sets and the regression lines between predicted and observed fire counts, are presented in Fig. 8.

5.3.1. Error histogram analysis

To analyse the distribution of errors in the predictions made by the AutoML-FIRE model on the testing dataset, we performed an error histogram analysis. The error was calculated as the difference between the observed and predicted fire counts. Histograms with 35 bins displaying these errors were plotted for all clusters in Fig. 9. Values to the right of the red line represent errors due to overestimation, while those to the left represent underestimation. Gaussian curves were fitted to the histograms to better understand the error distribution. The peak of the Gaussian curve aligns with the zero line, indicating an even distribution of errors around zero. This suggests that the model does not exhibit any significant bias in predicting forest fire occurrences.

5.3.2. Residual analysis

Residuals are the difference between the predictions of the model and the values of the regression line fitted on these predictions. Residual analysis is done to get a visual representation of the goodness of fit of our model. The residuals of the predictions of AutoML-FIRE on the testing dataset of all clusters are plotted in Fig. 10. The residuals are roughly uniformly distributed above and below the zero line, which proves the stochastic nature of the residuals. This confirms that the model does not have any inherent bias, and it indicates the generalisation capability of our model.

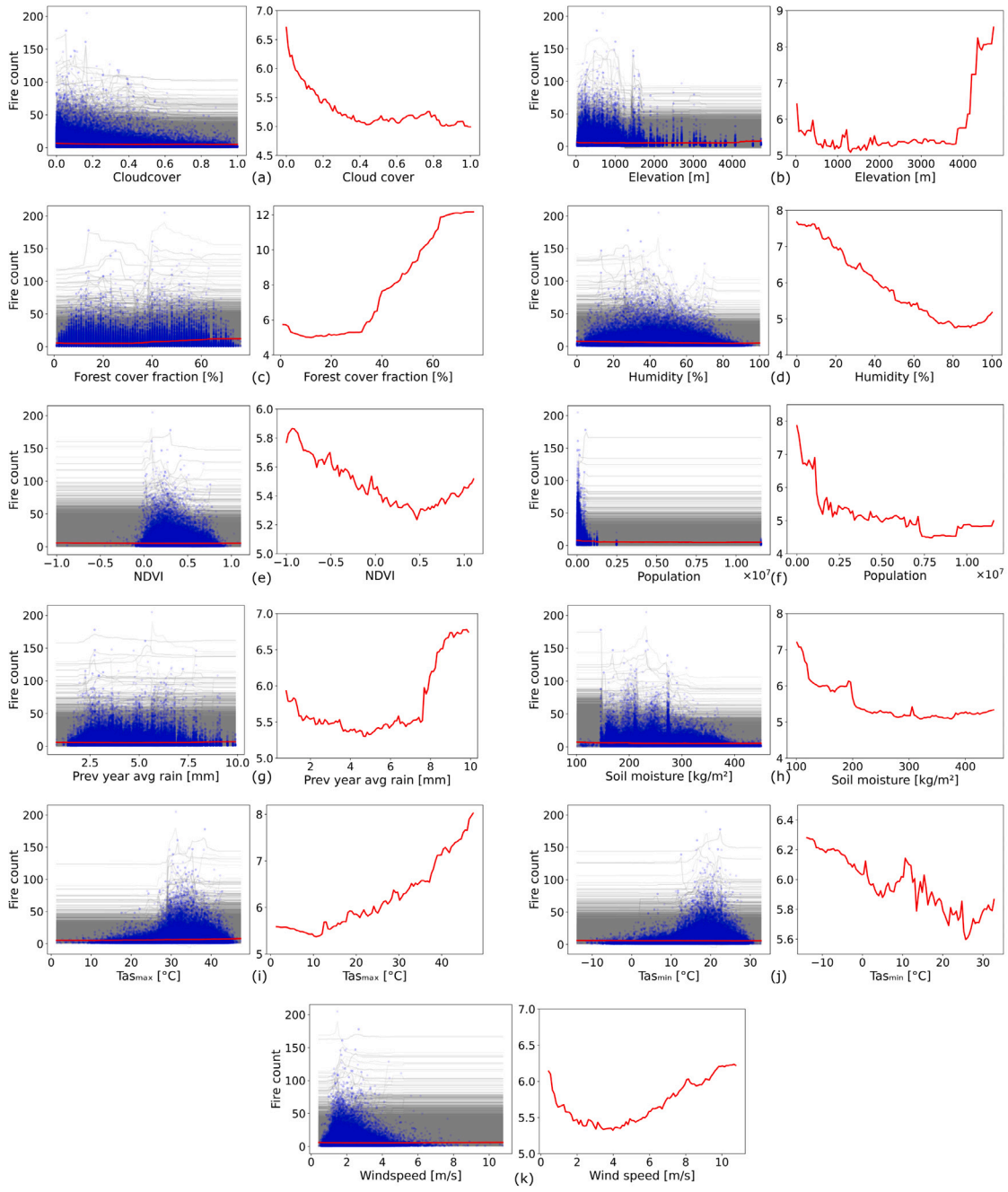


Fig. 6. Feature sensitivity analysis of the input features for the top 30% grids dataset to understand the trend across India. In the left plot of each subplot, the grey line illustrates the ICE curves, the blue dots represent the actual data points used to determine the ICE curves, and the red line represents the PDP line. The right side is the zoomed version of the PDP line plotted to observe the trend.

6. Discussion

6.1. Comparison with benchmark algorithms

We test the performance of the AutoML-Fire model with ten widely used machine learning and deep learning algorithms; long short-term memory (LSTM), linear regression (LR), extreme gradient boosting (XGBoost), decision tree (DT), K-nearest neighbour (KNN), artificial neural networks (ANN), elastic net, Bayesian regression (BayesReg), random forest (RF), polynomial regression (PolyReg), and generalised additive model (GAM). The algorithms were evaluated based on three key metrics: R, RMSE, and Bias. Table 3 summarises the comparative results across all clusters.

The results indicate that AutoML-FIRE consistently outperformed all other models in terms of RMSE across every cluster. Although it exhibited a slightly lower R in the Center & East cluster and in 30% of the grids, the difference was marginal. However, AutoML-FIRE exhibited a relatively higher bias compared to the other algorithms in all clusters. Despite this, the bias remained below 1.5, which is within an acceptable range given the model's exceptional performance in terms of RMSE and R. These results confirm AutoML-FIRE's overall effectiveness, particularly in achieving superior predictive accuracy across multiple metrics.

Analysis of variance (ANOVA) was conducted on the prediction errors across all benchmarking algorithms, including AutoML-FIRE, to assess the statistical significance of performance differences among

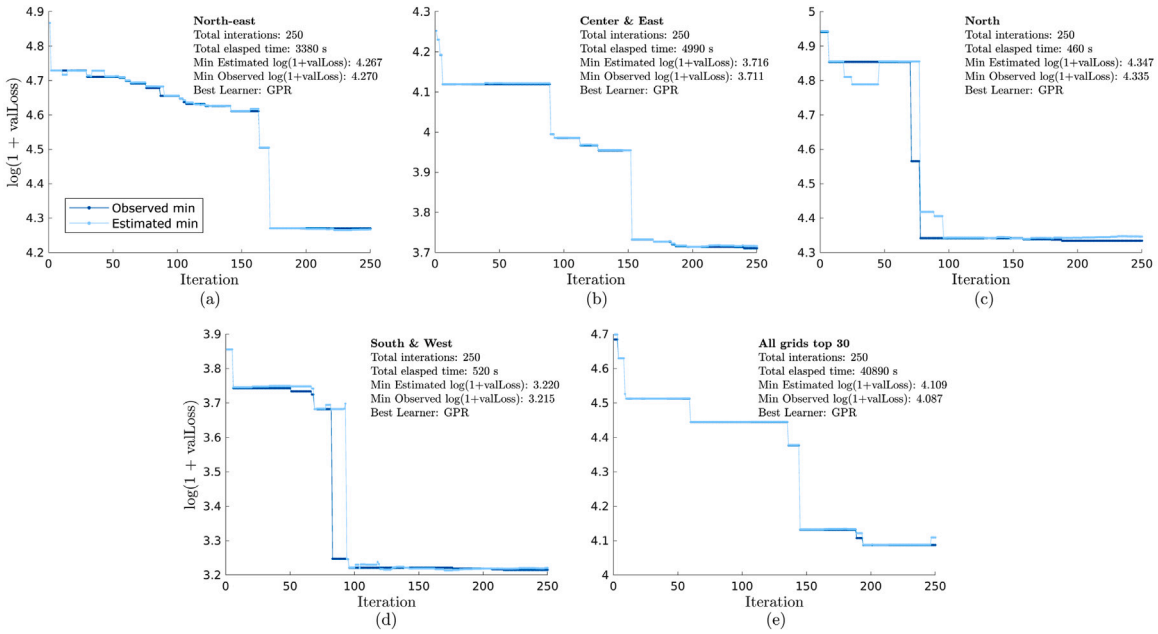


Fig. 7. Optimisation curve for each cluster depicting the minimum observed and estimated loss as the Bayesian optimisation progresses.

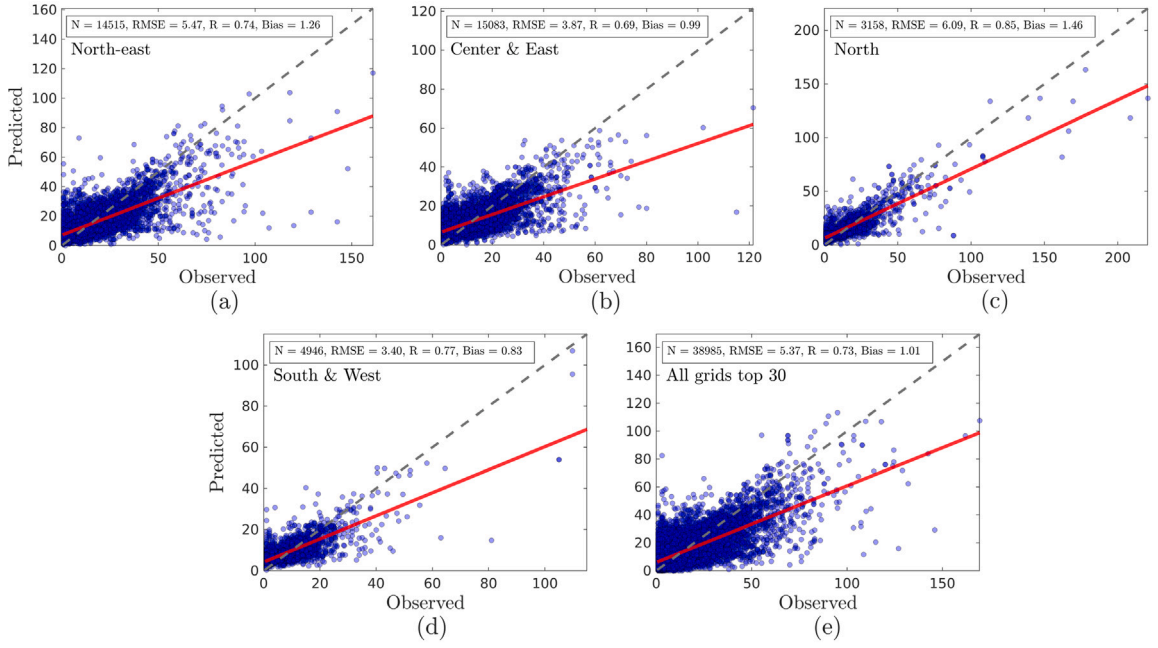


Fig. 8. Evaluation of AutoML-FIRE on testing data. In each subplot, the blue dots illustrate the predicted value vs. the observed value and the red line is the regression line fitted to this data. The grey dashed line shows the 1:1 line. The box displays the number of data points in the test set and the values of R, RMSE, and Bias.

the models. The results of the two-way ANOVA test are provided in the supplementary material, and those of one-way ANOVA test are illustrated in Fig. 11. It demonstrates that KNN, PolyReg, and GAM exhibit statistically significant variations in performance across all clusters, with the notable exception of the North cluster. In the North cluster, no statistically significant differences were observed among the benchmarking models, including AutoML-FIRE, suggesting comparable performance in this region. However, when analysing data from the top 30% of grids, AutoML-FIRE shows a significant divergence in performance compared to most of the other models. The models that deviate from AutoML-FIRE tend to cluster around zero error, indicating that while these models perform similarly to each other, AutoML-FIRE achieves distinct and superior outcomes in the top-performing grids.

6.2. Uncertainty analysis of the AutoML-FIRE

The reliability of a model can be assessed by evaluating its ability to manage randomness and variability in input data. To evaluate the robustness of our model, we performed an uncertainty analysis. In this analysis, we introduced uncertainties of ±5% and ±10% in each feature individually for 50% of the data points, while keeping the remaining dataset constant. AutoML-FIRE was then employed to predict outcomes on this modified dataset, and the percentage change in forest fire predictions relative to the original dataset was recorded. The results, presented in Fig. 12, show that the uncertainty in fire predictions across all clusters ranges from −2.87% to 1.82%. These minimal variations demonstrate the model's high reliability. The colour bar scale for

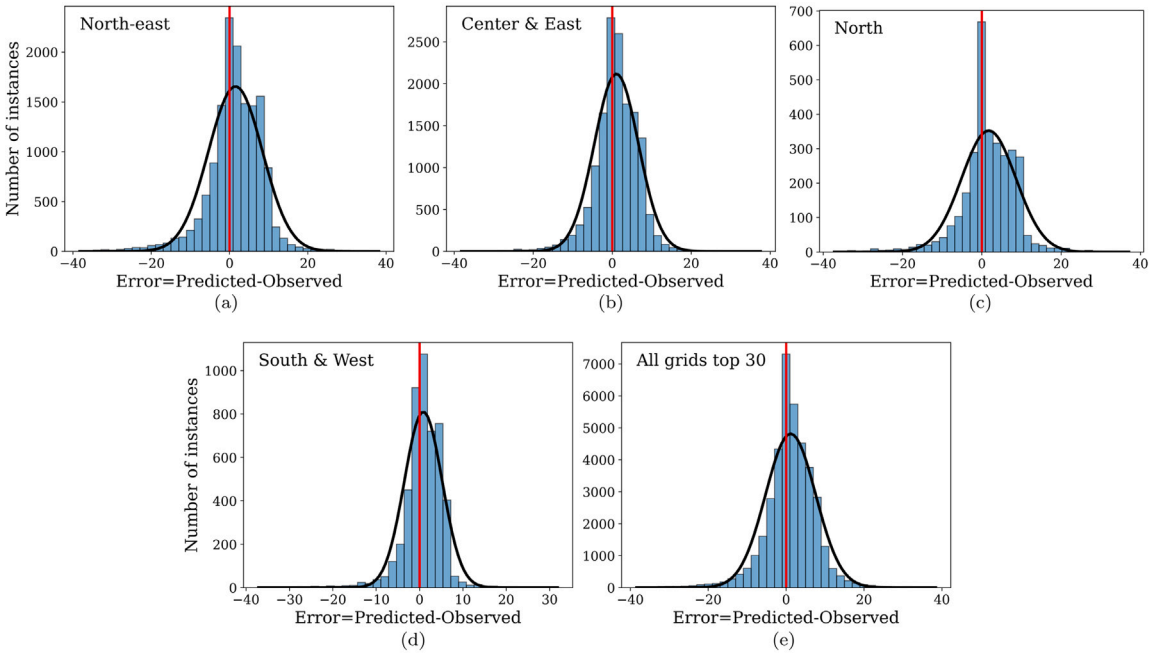


Fig. 9. Histograms of error in predictions of AutoML-FIRE on all clusters. In each sub-plot, the histogram has 35 bins, the black line illustrates the best-fit Gaussian curve on the histogram, and the red line indicates the zero error.

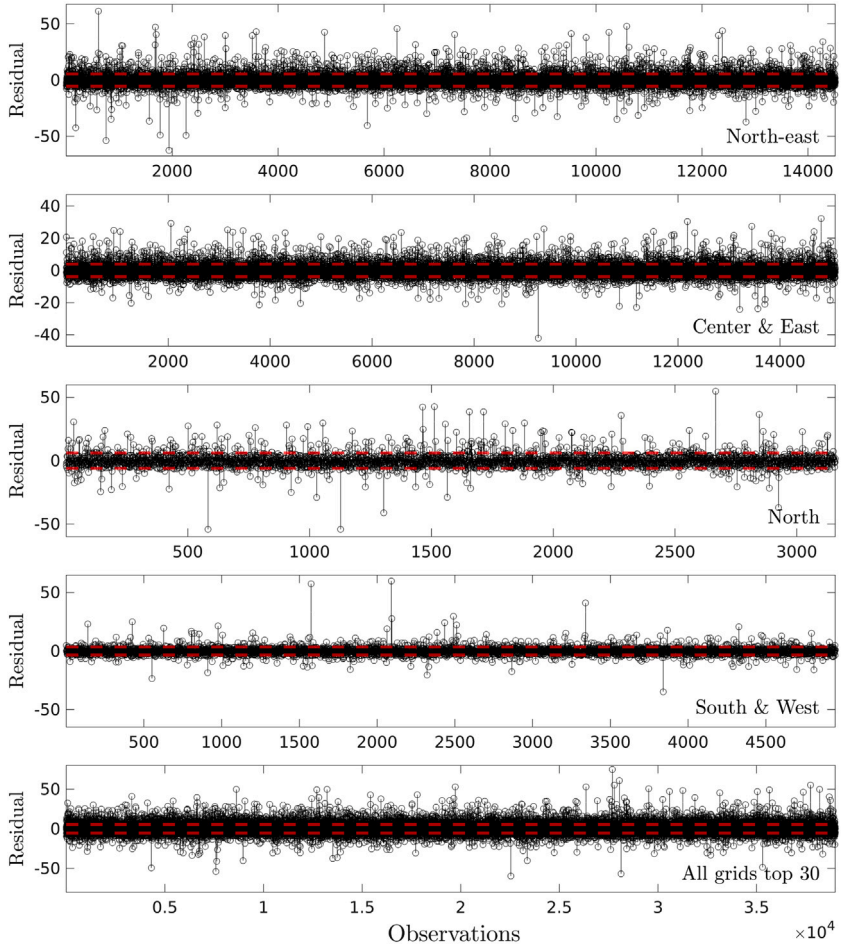


Fig. 10. Residuals indicating the difference between the predictions and the fitted regression line. The red dashed line represents the positive and negative RMSE values as a reference to determine the spread of residuals.

Table 3
Comparison with the benchmark algorithm for each cluster.

Cluster	Metrics	LSTM	LR	XGBoost	DT	KNN	ANN	Elasticnet	BayesReg	RF	PolyReg	GAM	AutoML-FIRE
North-east	R	0.5	0.39	0.73	0.44	0.66	0.5	0.39	0.39	0.71	0.48	0.43	0.74
	RMSE	10.56	11.24	8.41	11.01	9.19	10.59	11.24	11.24	8.66	10.71	11.06	5.47
	Bias	−0.25	0	0.27	0.05	1.15	0.38	0	0	0.59	0.01	−0.02	1.26
Centre & East	R	0.41	0.23	0.72	0.39	0.56	0.42	0.23	0.23	0.7	0.34	0.31	0.69
	RMSE	7.62	8.14	5.76	7.69	6.92	7.57	8.14	8.14	5.97	7.86	7.96	3.87
	Bias	0.13	−0.04	0.34	−0.03	1.16	−0.48	−0.04	−0.04	0.56	−0.02	0	0.99
North	R	0.48	0.31	0.8	0.62	0.71	0.53	0.31	0.31	0.78	0.46	0.47	0.85
	RMSE	12.5	13.53	8.46	11.22	10.01	12.05	13.53	13.53	8.84	12.66	12.54	6.09
	Bias	0.12	0.19	0.52	0.48	1.27	0.49	0.19	0.2	0.77	0.23	0.18	1.46
South &West	R	0.51	0.25	0.77	0.5	0.66	0.55	0.25	0.25	0.75	0.34	0.35	0.77
	RMSE	6.23	7.04	4.64	6.29	5.43	6.06	7.04	7.04	4.82	6.83	6.79	3.40
	Bias	−0.19	−0.03	0.35	0.22	0.96	−0.38	−0.03	−0.03	0.46	−0.01	−0.03	0.83
All grids top 30	R	0.47	0.28	0.75	0.44	0.65	0.47	0.28	0.28	0.74	0.4	0.35	0.73
	RMSE	9.49	10.32	7.13	9.63	8.16	9.48	10.32	10.32	7.18	9.85	10.04	5.37
	Bias	−0.13	0.01	0.22	0	1.24	0.67	0.01	0.01	0.6	−0.01	−0.01	1.01

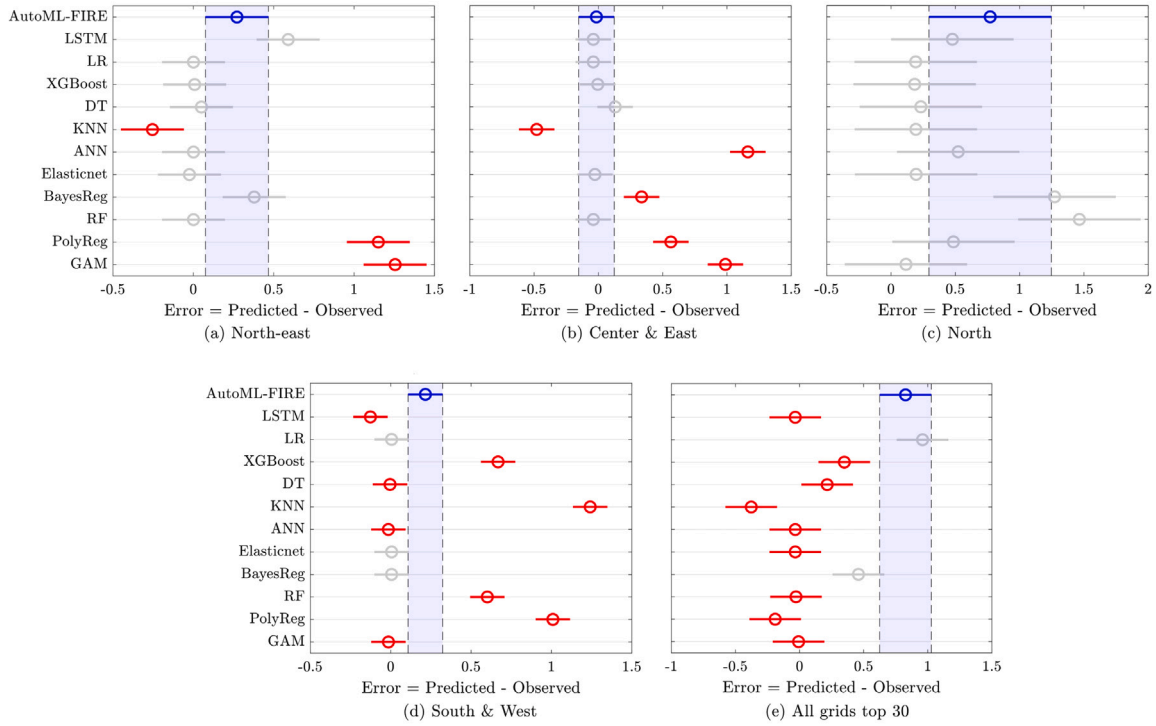


Fig. 11. Comparison of benchmarking algorithms with AutoML-FIRE based on errors in predictions using one-way ANOVA test. In each of the subplots, the blue bar along with the dashed line represents the comparison interval for the mean of the error of AutoML-FIRE. The grey line represents the model errors that overlap with the comparison interval and thus are statistically similar to AutoML-FIRE. The red line represents statistically different model results.

the individual clusters provides insight into the range of variation within each cluster. From this, we can conclude that the South & West cluster is the least susceptible to uncertainty. However, the model is most sensitive to uncertainty in the previous year's average rainfall (PYR) and forest cover fraction among the variables across the clusters. Sensitivity to forest cover emphasises the need for enhanced mapping accuracy of forested areas. Uncertainties in forest cover fraction can introduce variability in fire count predictions, potentially resulting in the misallocation of resources by fire management authorities. By including uncertainty intervals, decision-makers are better equipped to handle system reliability under varying conditions and can incorporate these findings into risk management frameworks that account for potential variability in predictions, thereby improving the overall robustness of wildfire management systems. Overall, this uncertainty analysis underscores the stability and robustness of AutoML-FIRE in handling variability within the dataset.

6.3. Spatial distribution analysis

To demonstrate that the model's superior performance was not simply due to a favourable train-test split, and to confirm that the model is independent of the spatial variability of the training set, we conducted a spatial distribution analysis. This is essential because models are trained on data from certain regions might perform better purely due to the characteristics of that region, which would undermine the ability of the model to generalise. To address this, we trained and evaluated the model using 30 different seed values, resulting in 30 distinct random splits of the dataset into training and testing data. The evaluation metrics for both the training and testing datasets are provided in the supplementary material, with the $\mu \pm \sigma$ reported in Table 4. This result can be used to conclude that the model's performance is independent of the spatial distribution of the data, as the deviations from the mean are minimal in all cases. This spatial independence is crucial as it validates the generalisation capability of the model across

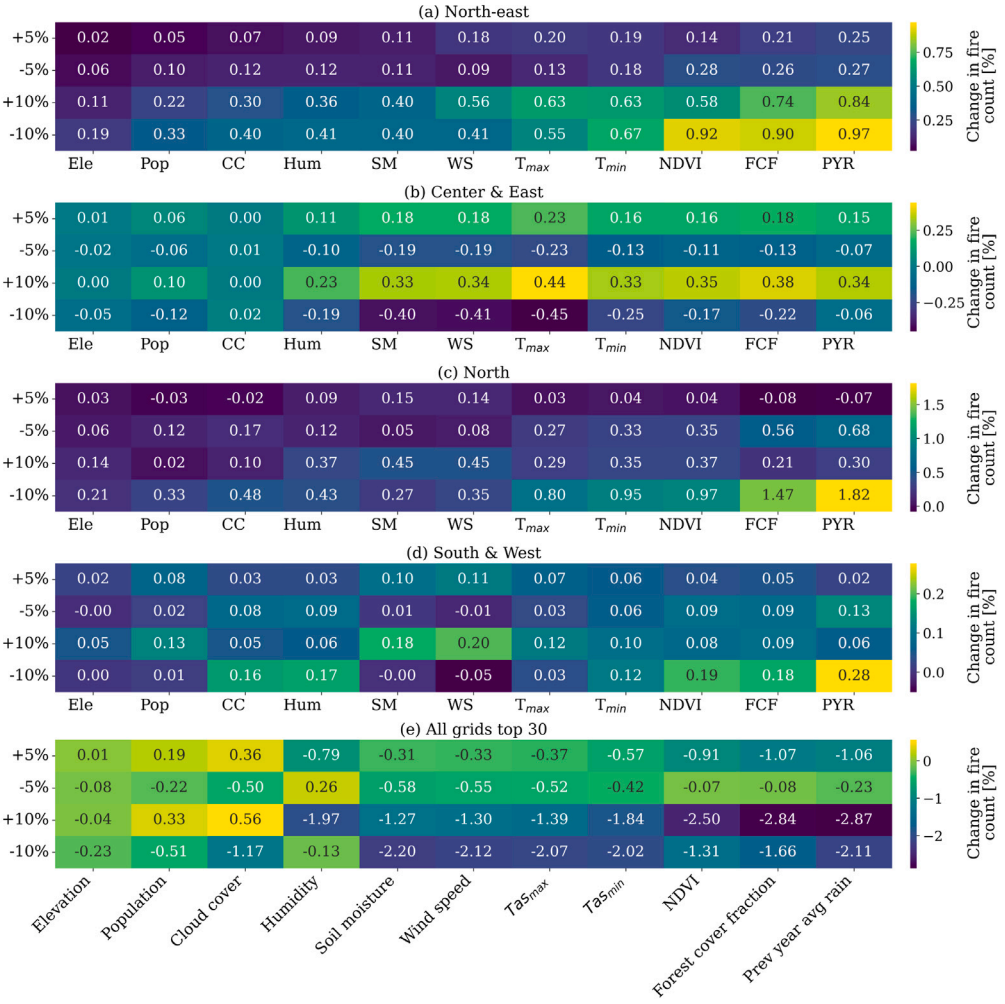


Fig. 12. The figure depicts feature sensitivity by applying ±5% and ±10% uncertainty to each feature. It demonstrates how perturbations in feature values influence the model's predictions in %.

Table 4
Spatial distribution analysis.

Cluster	Training			Testing		
	R	RMSE	Bias	R	RMSE	Bias
North-east	0.98 ± 0.01	2.13 ± 0.30	0.08 ± 0.05	0.74 ± 0.01	5.78 ± 0.20	1.01 ± 0.21
Center & East	0.98 ± 0.01	1.42 ± 0.16	0.06 ± 0.09	0.70 ± 0.01	3.97 ± 0.15	0.87 ± 0.23
North	0.98 ± 0.01	2.31 ± 0.27	0.05 ± 0.13	0.82 ± 0.03	6.05 ± 0.40	1.39 ± 0.41
South & West	0.99 ± 0.00	1.03 ± 0.12	0.04 ± 0.02	0.76 ± 0.05	3.40 ± 0.20	0.75 ± 0.18
All grids top 30	0.96 ± 0.05	2.31 ± 0.63	0.05 ± 0.09	0.72 ± 0.02	5.19 ± 0.27	0.80 ± 0.31

different regions and ensures that its performance is not confined to any particular geographical area or specific training dataset. Therefore, our model is robust and reliable when applied to diverse spatial contexts, enhancing its applicability in real-world scenarios.

6.4. Ablation analysis

We performed an ablation analysis of our model to assess how its performance is impacted by restricting the number of features. AutoML-FIRE was trained and evaluated using nine different combinations of features that demonstrated high importance across the clusters. The evaluation metrics R, RMSE, and Bias for each cluster and feature combination are reported in Table 5. As expected, due to the high relevance of these features, their combinations are able to predict a significant portion of the target variable, as evidenced by the strong R values. Remarkably, the model maintains high accuracy even when the number of features is reduced to just three.

A notable finding from the analysis is that AutoML-FIRE consistently performs much better in the North cluster across nearly all feature combinations, except when only the population feature is used. The population feature exhibits anomalous behaviour in this region and fails to capture the trend in the North cluster, as discussed in Section 5.1. This anomaly is likely attributed to unique socio-environmental factors specific to the North cluster. On the other hand, the model's performance significantly improves when all features are combined, reinforcing the conclusion that a complex interplay of multiple factors drives forest fires. This underscores the importance of considering a holistic set of variables when building predictive models for such intricate natural phenomena.

Table 5
Ablation analysis of the input features.

Considered features	Metrics	North-east	Center & East	North	South & West	All grids top 30
Population	R	0.43	0.40	0.39	0.40	0.44
	RMSE	4.88	3.2	5.33	2.65	4.33
	Bias	0.08	0.08	−0.25	−0.1	0.03
Population + Humidity	R	0.46	0.43	0.65	0.40	0.43
	RMSE	4.48	2.78	6.08	3.04	3.24
	Bias	0.09	0.09	−0.20	0.07	0.02
Soil moisture	R	0.33	0.31	0.63	0.40	0.36
	RMSE	4.00	3.56	7.48	3.07	4.87
	Bias	0.09	0.26	0.77	−0.08	0.44
Soil moisture + Elevation	R	0.47	0.41	0.51	0.49	0.44
	RMSE	4.69	2.61	5.00	2.93	3.39
	Bias	0.06	0.08	−0.11	−0.11	0.01
Soil moisture + Tas_{max}	R	0.40	0.39	0.66	0.4	0.32
	RMSE	4.32	3.05	6.84	3.26	6.39
	Bias	0.1	0.21	0.26	0.11	0.72
Soil moisture + Tas_{max} + Humidity	R	0.42	0.39	0.66	0.46	0.46
	RMSE	4.06	2.93	6.59	3.65	4.48
	Bias	0.78	0.06	−0.16	0.45	0.2
Soil moisture + Humidity + Population	R	0.47	0.46	0.68	0.53	0.46
	RMSE	4.72	2.92	6.35	3.06	4.5
	Bias	0.09	0.09	−0.01	−0.12	0.13
Humidity + Forest cover fraction + Population	R	0.47	0.48	0.63	0.55	0.46
	RMSE	4.42	3.03	5.98	2.98	3.36
	Bias	0.07	0.08	−0.27	−0.09	0.02
Elevation + Population + Humidity	R	0.46	0.46	0.66	0.53	0.45
	RMSE	4.73	3.08	6.42	2.83	3.62
	Bias	0.07	0.08	−0.23	−0.13	0.02
All features	R	0.74	0.69	0.85	0.77	0.73
	RMSE	5.47	3.87	6.09	3.40	5.37
	Bias	1.26	0.99	1.46	0.83	0.01

6.5. Impact analysis of the AutoML-FIRE model

In this section, we examine the potential impact of AutoML-FIRE on the field of forest fire prediction. We analyse the environmental, social, and economic benefits of our model and its role in aiding the decision-making processes of authorities responsible for forest fire management.

- **Quantifying forest fire**: Traditionally, forest fire prediction has been approached as a classification problem. In this study, we quantified forest fires by predicting the forest fire count using our model. Quantifying forest fires provided a more precise assessment of risk, allowing for more targeted precautionary measures.
- **Pan-India and regional study**: To the best of our knowledge, this is the first pan-India study of forest fires. This provides a comprehensive overview of forest fire occurrences and their dependencies on various factors across India's diverse geographical regions. Additionally, by clustering the data into sub-regions, we achieved localised yet improved predictive accuracy.
- **Improved accuracy**: Through the evaluation of AutoML-FIRE and comparison with benchmarking algorithms, we demonstrated the superior performance and predictive capabilities of our model. As a result, our forest fire predictions showed marked improvements, minimising the likelihood of false alarms.
- **Robust and stable nature of the model**: The various analyses conducted established the robustness and stability of AutoML-FIRE. These results confirmed that the model is reliable for forest fire prediction, making it suitable for integration into warning systems.
- **Broader adaption**: With the increase in forest fires worldwide, the "AutoML-FIRE" framework offers a scalable and adaptable approach to improve regional predictive capabilities. By incorporating region-specific data, it can enhance the accuracy of fire risk assessments across diverse geographic contexts.
- **Role in decision-making**: AutoML-FIRE has the potential to assist authorities in forest fire management by providing accurate predictions, which would aid in more effective mitigation and response strategies. It could be particularly useful for national governments in fairly allocating resources to states that are more vulnerable to forest fires. Furthermore, local authorities could use these predictions to better prepare for impending fire threats.
- **Societal impact**: One of the most significant impacts of forest fire prediction is the ability to enable local communities to take necessary precautions and prepare in advance. By incorporating this model into early warning systems, the loss of human life and economic damage could be minimised.

6.6. Limitations and future work

AutoML-FIRE represents a significant advancement in the domain of forest fire prediction, demonstrating a high degree of accuracy and reliability. However, despite its strengths, the model has certain limitations that could be addressed in future work.

- **Number of variables:** Forest fires result from a complex interplay of various factors. Although we considered a broad set of input variables, there may be additional factors influencing forest fire occurrences. In particular, future work could incorporate more variables that account for anthropogenic influences, such as land use changes, deforestation, and human activity in vulnerable regions.
- **Reliance on historical data:**
  One of the key limitations of this study is that the dataset used ends in 2018, which may affect the current relevance of the predictions given recent changes in climate patterns, land use, and human activity. Forest fire dynamics are influenced by both short-term variability and long-term trends, and events post-2018, such as increasingly frequent extreme heatwaves or land-use change,

are not captured in the training data. As a result, model performance in recent years may differ, and retraining with updated datasets could be necessary to maintain accuracy. Incorporating more recent data, as it becomes available, will be essential for improving model robustness and ensuring applicability to current fire risk scenarios.

- **Lack of real-time processing:** The current version of AutoML-FIRE lacks the ability to process real-time data, which limits its use in emergency situations that require timely and dynamic predictions. In such scenarios, access to real-time environmental variables such as temperature, wind speed, and humidity is essential for accurate forecasting. Integrating AutoML-FIRE into operational workflows and real-time data-sharing platforms could enable continuous updates, making the model more responsive and suitable for real-world fire monitoring and early warning systems.
- **Spatial resolution:** In this study, fire counts were predicted for a grid size of $0.25° \times 0.25°$. With the availability of higher-resolution input variables, AutoML-FIRE could be deployed to predict forest fires at a finer spatial resolution. This would enable a more targeted and region-specific approach to fire management and mitigation efforts.
- **Nowcasting capabilities:** AutoML-FIRE can predict forest fires in near real-time when provided with up-to-date data. However, a key direction for future research is the development of models that can forecast forest fires well in advance. Such advancements would improve preparedness and response strategies, potentially reducing both human and environmental losses.
- **Extreme weather conditions:** Our study does not fully account for extreme weather conditions, such as sudden wind shifts or prolonged droughts, which could significantly alter fire behaviour and spread.
- **Post-fire effects:** While this study focuses on forest fire prediction, it does not analyse the large-scale environmental impacts that occur post-fire. Future research could leverage machine learning models to assess post-fire effects, such as changes in soil quality and variations in air pollution levels, particularly concentrations of particulate matter ($PM_{2.5}$ and PM_{10}). Developing predictive frameworks for these factors would enhance our understanding of fire-induced ecological changes and aid in post-fire management strategies.

7. Conclusion

Our findings demonstrate that AutoML-FIRE provides an accurate and reliable approach for predicting forest fire occurrences across high-risk regions in India. The model consistently surpasses conventional benchmarks in predictive performance, with robust results validated through spatial distribution and uncertainty analyses. These results underscore the potential of AutoML-FIRE as an effective tool for operational early warning systems, offering critical support to local authorities in mitigating the environmental and socio-economic impacts of forest fires. Integrating AutoML-FIRE into existing frameworks could significantly enhance preparedness and resilience against escalating fire risks.

CRediT authorship contribution statement

Saurabh Toraskar: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation. **Adil Khan:** Writing – original draft, Validation, Data curation. **M. Niranjannaik:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Abhilash Singh:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Kumar Gaurav:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Conceptualization.

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Software and data availability

- Software name: AutoML-FIRE (**Auto**mated **M**achine-**L**earning approach to predict forest **FIRE**)
- Developer: Toraskar et al.
- Contact information: kgaurav@iiserb.ac.in
- First year available: 2025
- Program language: MATLAB
- Cost: Free
- Software availability: https://abhilashsingh.net/codes.html
- Data availability:
  - Input features: The input features utilised in this study are sourced from publicly available databases and can be downloaded from the respective websites as explained in Section 4.1
  - Response variable: We have downloaded the forest fire data from the Forest Survey of India (FSI) website (https://fsiforestfire.gov.in/index.php). The authors do not have permission to share this data.
- Infrastructure used: NVIDIA A40 GPU with 46 GB VRAM and an Intel Xeon Gold 6338 CPU (64 cores, 128 threads)

Declaration of competing interest

We have no conflicts of interest to disclose.

Acknowledgements

We acknowledge IISER Bhopal for providing institutional support. ST acknowledges the Department of Science and Technology (DST), Government of India, for INSPIRE-SHE (Scholarship for Higher Education). The authors sincerely appreciate the thoughtful feedback and helpful suggestions provided by the editor and the three anonymous reviewers.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.envsoft.2025.106578.

References

Adler, R., Wang, J.-J., Sapiano, M., Huffman, G., Bolvin, D., Nelkin, E., et al., 2020. Global precipitation climatology project (GPCP) climate data record (CDR), version 1.3 (daily).

Ahmad, F., Goparaju, L., 2019. Forest fire trend and influence of climate variability in India: A geospatial analysis at national and local scale. Ekol. (Bratislava) 38 (1), 49–68. http://dx.doi.org/10.2478/eko-2019-0005.

Alex Goldstein, J.B., Pitkin, E., 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J. Comput. Graph. Statist. 24 (1), 44–65. http://dx.doi.org/10.1080/10618600.2014.907095, arXiv:https://doi.org/10.1080/10618600.2014.907095.

Attri, V., Dhiman, R., Sarvade, S., 2020. A review on status, implications and recent trends of forest fire management. Arch. Agric. Environ. Sci. 5 (4), 592–602.

Babu, K.N., Gour, R., Ayushi, K., Ayyappan, N., Parthasarathy, N., 2023. Environmental drivers and spatial prediction of forest fires in the western ghats biodiversity hotspot, India: An ensemble machine learning approach. Forest Ecol. Manag. 540, 121057.

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci. Data 5 (1), 1–12.

Beer, T., 1991. The interaction of wind and fire. Bound.-Layer Meteorol. 54 (3), 287–308.

Bhadani, V., Singh, A., Kumar, V., Gaurav, K., 2024. Nature-inspired optimal tuning of input membership functions of fuzzy inference system for groundwater level prediction. Environ. Model. Softw. 175, 105995.

Boer, M.M., Resco de Dios, V., Bradstock, R.A., 2020. Unprecedented burn area of Australian mega forest fires. Nat. Clim. Chang. 10 (3), 171–172.

Boogaard, H., Schubert, J., De Wit, A., Lazebnik, J., Hutjes, R., Van der Grijn, G., et al., 2020. Agrometeorological indicators from 1979 to present derived from reanalysis. Copernic. Clim. Chang. Serv. (C3S) Clim. Data Store (CDS) 10.

Branco, P., Ribeiro, R.P., Torgo, L., 2016. UBL: an R package for utility-based learning. arXiv preprint arXiv:1604.08079.

Branco, P., Torgo, L., Ribeiro, R.P., 2017. SMOGN: a pre-processing approach for imbalanced regression. In: First International Workshop on Learning with Imbalanced Domains: Theory and Applications. PMLR, pp. 36–50.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., 2017. Classification and Regression Trees. Routledge.

Brotak, E.A., 1991. Low-level temperature, moisture and wind profiles preceding major wildland fires. In: Proceedings 11th Conference on Fire and Forest Meteorology, Missoula. pp. 503–510.

Center for International Earth Science Information Network-CIESIN-Columbia University, 2018. Gridded population of the world, version 4 (GPWv4): Population density, revision 11.

Chaparro, D., Vall-llossera, M., Piles, M., Camps, A., Rüdiger, C., 2015. Low soil moisture and high temperatures as indicators for forest fire occurrence and extent across the Iberian Peninsula. In: 2015 IEEE International Geoscience and Remote Sensing Symposium. IGARSS, pp. 3325–3328. http://dx.doi.org/10.1109/IGARSS.2015.7326530.

Chuvieco, E., Cocero, D., Riano, D., Martin, P., Martınez-Vega, J., de la Riva, J., Perez, F., 2004. Combining NDVI and surface temperature for the estimation of live fuel moisture content in forest fire danger rating. Remote Sens. Environ. 92 (3), 322–331. http://dx.doi.org/10.1016/j.rse.2004.01.019, URL: https://www.sciencedirect.com/science/article/pii/S0034425704001531, Forest Fire Prevention and Assessment.

Copernicus Climate Change Service, Climate Data Store, 2021. Temperature and Precipitation Gridded Data for Global and Regional Domains Derived from In-Situ and Satellite Observations.

Cortez, P., Morais, A., 2007a. New trends in artificial intelligence. In: Proceedings of the 13th Portuguese Conference on Artificial Intelligence (EPIA 2007), GuimarÃes, Portugal. pp. 3–7.

Cortez, P., Morais, A.d.R., 2007b. A data mining approach to predict forest fires using meteorological data.

Dhar, T., Bhatta, B., Aravindan, S., 2023. Forest fire occurrence, distribution and risk mapping using geoinformation technology: A case study in the sub-tropical forest of the Meghalaya, India. Remote. Sens. Appl.: Soc. Environ. 29, 100883. http://dx.doi.org/10.1016/j.rsase.2022.100883, URL: https://www.sciencedirect.com/science/article/pii/S2352938522001914.

DiMiceli, C., Sohlberg, R., Townshend, J., 2022. MODIS/Terra vegetation continuous fields yearly L3 global 250 m SIN grid V061. NASA EOSDIS Land Process. DAAC.

Earth Resources Observation and Science Center, U.S. Geological Survey, U.S. Department of the Interior, 1997. USGS 30 ARC-second global elevation data, GTOPO30. URL: https://doi.org/10.5065/A1Z4-EE71.

Friedman, J.H., 2000. Greedy function approximation: A gradient boosting machine. Ann. Stat. 29, 1189–1232.

Gholami, S., Kodandapani, N., Wang, J., Ferres, J.L., 2021. Where there's smoke, there's fire: Wildfire risk predictive modeling via historical climate data. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35, pp. 15309–15315.

Hainmueller, J., Hazlett, C., 2014. Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. Political Anal. 22 (2), 143–168. http://dx.doi.org/10.1093/pan/mpt019.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, vol. 2, Springer.

Jain, P., Castellanos-Acuna, D., Coogan, S.C., Abatzoglou, J.T., Flannigan, M.D., 2022. Observed increases in extreme fire weather driven by atmospheric humidity and temperature. Nat. Clim. Chang. 12 (1), 63–70.

Kang, Y., Jang, E., Im, J., Kwon, C., Kim, S., 2020. Developing a new hourly forest fire risk index based on catboost in South Korea. Appl. Sci. 10 (22), 8213.

Kim, S.J., Lim, C.-H., Kim, G.S., Lee, J., Geiger, T., Rahmati, O., Son, Y., Lee, W.-K., 2019. Multi-temporal analysis of forest fire probability using socio-economic and environmental variables. Remote. Sens. 11 (1), 86.

Kong, B., 2024. A comparative analysis of machine learning models for wildfire prediction.

Kumar, A., Gaurav, K., Singh, A., Yaseen, Z.M., 2024. Assessment of machine learning models to predict daily streamflow in a semiarid river catchment. Neural Comput. Appl. 36 (21), 13087–13106.

Kunz, N., 2020. SMOGN: Synthetic minority over-sampling technique for regression with Gaussian noise. URL: https://pypi.org/project/smogn/.

Loh, W.-Y., Shih, Y.-S., 1997. Split selection methods for classification trees. Statist. Sinica 815–840.

Lundberg, S.M., Erion, G.G., Lee, S.-I., 2018. Consistent individualized feature attribution for tree ensembles. ArXiv abs/1802.03888, URL: https://api.semanticscholar.org/CorpusID:3626364.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS '17, Curran Associates Inc., Red Hook, NY, USA, pp. 4768–4777.

Maselli, F., Romanelli, S., Bottai, L., Zipoli, G., 2003. Use of NOAA-AVHRR NDVI images for the estimation of dynamic fire risk in Mediterranean areas. Remote Sens. Environ. 86 (2), 187–197.

Mina, U., Dimri, A.P., Farswan, S., 2023. Forest fires and climate attributes interact in central Himalayas: an overview and assessment. Fire Ecol. 19, http://dx.doi.org/10.1186/s42408-023-00177-4.

Molnar, C., 2022. Interpretable Machine Learning, second ed. URL: https://christophm.github.io/interpretable-ml-book.

Pérez-Cabello, F., Cerdà, A., De La Riva, J., Echeverría, M., García-Martín, A., Ibarra, P., Lasanta, T., Montorio, R., Palacios, V., 2012. Micro-scale post-fire surface cover changes monitored using high spatial resolution photography in a semiarid environment: A useful tool in the study of post-fire soil erosion processes. J. Arid. Environ. 76, 88–96.

Pfister, G., McKenzie, R., Liley, J., Thomas, A., Forgan, B., Long, C.N., 2003. Cloud coverage based on all-sky imaging and its impact on surface solar irradiance. J. Appl. Meteorol. Clim. 42 (10), 1421–1434.

Radke, D., Hessler, A., Ellsworth, D., 2019. FireCast: Leveraging deep learning to predict wildfire spread.. In: IJCAI. pp. 4575–4581.

Rasmussen, C., Bousquet, O., Luxburg, U., Rätsch, G., 2004. Gaussian processes in machine learning. In: Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures. Vol. 3176, pp. 63–71. http://dx.doi.org/10.1007/978-3-540-28650-9_4.

Reddy, C.S., Bird, N.G., Sreelakshmi, S., Manikandan, T.M., Asra, M., Krishna, P.H., Jha, C., Rao, P., Diwakar, P., 2019. Identification and characterization of spatio-temporal hotspots of forest fires in South Asia. Environ. Monit. Assess. 191, 1–17.

Ribeiro, R.P., 2011. Utility-Based Regression (Ph. D. dissertation).

Robinne, F.-N., Secretariat, F., 2021. Impacts of disasters on forests, in particular forest fires. UNFFS Backgr. Pap..

Roy, A., Purohit, R., 2018. Indian subcontinent: Geomorphic and geophysical traits. Indian Shield. 2018, 13–30.

Saha, S., Bera, B., Shit, P.K., Bhattacharjee, S., Sengupta, N., 2023. Prediction of forest fire susceptibility applying machine and deep learning algorithms for conservation priorities of forest resources. Remote. Sens. Appl.: Soc. Environ. 29, 100917.

Sannigrahi, S., Pilla, F., Basu, B., Basu, A.S., Sarkar, K., Chakraborti, S., Joshi, P.K., Zhang, Q., Wang, Y., Bhatt, S., et al., 2020. Examining the effects of forest fire on terrestrial carbon emission and ecosystem production in India using remote sensing approaches. Sci. Total Environ. 725, 138331.

Schultz, M.G., Heil, A., Hoelzemann, J.J., Spessa, A., Thonicke, K., Goldammer, J.G., Held, A.C., Pereira, J.M., van Het Bolscher, M., 2008. Global wildland fire emissions from 1960 to 2000. Glob. Biogeochem. Cycles 22 (2).

Seibert, J., McDonnell, J.J., Woodsmith, R.D., 2010. Effects of wildfire on catchment runoff response: a modelling approach to detect changes in snow-dominated forested catchments. Hydrol. Res. 41 (5), 378–390.

Singh, A., Amutha, J., Nagar, J., Sharma, S., Lee, C.-C., 2022. AutoML-ID: Automated machine learning model for intrusion detection using wireless sensor network. Sci. Rep. 12 (1), 9074.

Singh, A., Gaurav, K., 2024. PIML-SM: Physics-informed machine learning to estimate surface soil moisture from multi-sensor satellite images by leveraging swarm intelligence. IEEE Trans. Geosci. Remote Sens..

Singh, A., Nagar, J., Sharma, S., Kotiyal, V., 2021. A Gaussian process regression approach to predict the k-barrier coverage probability for intrusion detection in wireless sensor networks. Expert Syst. Appl. 172, 114603.

Singh, A., Patel, S., Bhadani, V., Kumar, V., Gaurav, K., 2024. AutoML-GWL: Automated machine learning model for the prediction of groundwater level. Eng. Appl. Artif. Intell. 127, 107405.

Su, Y., Zhao, L., Li, H., Li, X., Chen, J., Ge, Y., 2024. An efficient task implementation modeling framework with multi-stage feature selection and AutoML: A case study in forest fire risk prediction. Remote. Sens. 16 (17), 3190.

Toledo, M., Poorter, L., Peña-Claros, M., Alarcón, A., Balcázar, J., Leaño, C., Licona, J.C., Llanque, O., Vroomans, V., Zuidema, P., Bongers, F., 2011. Climate is a stronger driver of tree and forest growth rates than soil and disturbance. J. Ecol. 99 (1), 254–264. http://dx.doi.org/10.1111/j.1365-2745.2010.01741.x.

Torgo, L., Branco, P., Ribeiro, R.P., Pfahringer, B., 2015. Resampling strategies for regression. Expert Syst. 32 (3), 465–476.

Torgo, L., Ribeiro, R., 2007. Utility-based regression. In: Knowledge Discovery in Databases: PKDD 2007: 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007. Proceedings 11. Springer, pp. 597–604.

Torgo, L., Ribeiro, R.P., Pfahringer, B., Branco, P., 2013. Smote for regression. In: Portuguese Conference on Artificial Intelligence. Springer, pp. 378–389.

Vapnik, V., Golowich, S., Smola, A., 1996. Support vector method for function approximation, regression estimation and signal processing. Adv. Neural Inf. Process. Syst. 9.

Vega-Garcia, C., Lee, B., Woodard, P., Titus, S., 1996. Applying neural network technology to human-caused wildfire occurrence prediction.

Venkatesh, K., Preethi, K., Ramesh, H., 2020. Evaluating the effects of forest fire on water balance using fire susceptibility maps. Ecol. Indic. 110, 105856.

Vermot, E., 2022. Noaa Climate Data Record (Cdr) of Avhrr Normalized Difference Vegetation Index (Ndvi), Version 5. NOAA National Centers for Environmental Information, http://dx.doi.org/10.7289/V5ZG6QH9www.ncei.noaa.gov/access/metadata/landing-page/bin/iso.

Wang, C.G., Zheng, X.B., Wang, A.Z., Dai, G.H., Zhu, B.K., Zhao, Y.M., Dong, S.J., Zu, W.Z., Wang, W., Zheng, Y.G., Li, J.G., Li, M.-H., 2021. Temperature and precipitation diversely control seasonal and annual dynamics of litterfall in a temperate mixed mature forest, revealed by long-term data analysis. J. Geophys. Res.: Biogeosciences 126 (7), e2020JG006204. http://dx.doi.org/10.1029/2020JG006204, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020JG006204, URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JG006204 e2020JG006204 2020JG006204.

Wijayanto, A.K., Sani, O., Kartika, N.D., Herdiyeni, Y., 2017. Classification model for forest fire hotspot occurrences prediction using ANFIS algorithm. In: IOP Conference Series: Earth and Environmental Science. Vol. 54, IOP Publishing, 012059.

Xu, Y., Li, D., Ma, H., Lin, R., Zhang, F., 2022. Modeling forest fire spread using machine learning-based cellular automata in a GIS environment. Forests 13 (12), 1974.

Yang, S., Lupascu, M., Meel, K.S., 2021. Predicting forest fire using remote sensing data and machine learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35, pp. 14983–14990.

Zhang, S., Pan, M., 2024. An AutoML-Powered analysis framework for forest fire forecasting: Adapting to climate change dynamics. Atmosphere 15 (12), 1481.