

BUILDING A MODEL OF EARLY FIRE DETECTION USING MACHINE LEARNING METHODS AND SATELLITE IMAGERY

Oleksii Basmanov

Corresponding author

Doctor of Technical Sciences, Professor, Leading Researcher
Scientific and Testing Department of Fire Protection and Fire Extinguishing Systems Research
of the Scientific and Research Center of Research and Testing*

E-mail: basmanov_oleksii@nuczu.edu.ua

ORCID: <https://orcid.org/0000-0002-6434-6575>

Serhii Ohurtsov

Candidate of Technical Sciences, Senior Researcher*

ORCID: <https://orcid.org/0000-0003-3224-9436>

Roman Ponomarenko

Doctor of Technical Sciences, Professor, Head of Institute*

ORCID: <https://orcid.org/0000-0002-6300-3108>

Andrii Khyzhniak

Doctor of Philosophy (PhD), Assistant Professor, Head**

ORCID: <https://orcid.org/0009-0003-5518-0723>

Oleksandr Lysenko

ORCID: <https://orcid.org/0009-0006-2720-0266>

Yevhen Ryzhkov

Leading Specialist

Telecommunications, Information Technology and 112 System Sector**

ORCID: <https://orcid.org/0009-0003-6470-2159>

Dmytro Chalyy

Leading Specialist

Department of Civil Protection and Preventive Activities***

ORCID: <https://orcid.org/0000-0002-7136-6582>

Iryna Chala

Leading Inspector

Sector for Organization of Security Measures for Critical Infrastructure

Department of Civil Protection and Preventive Activities***

ORCID: <https://orcid.org/0009-0009-2221-7503>

Anastasiia Hryshchenko

Lecturer

Department of State Supervision in Fire and Technogenic Safety

National University of Civil Protection of Ukraine

Onopriyenko str., 8, Cherkasy, Ukraine, 18034

ORCID: <https://orcid.org/0009-0003-1035-2624>

*Institute of Scientific Research on Civil Protection

of the National University of Civil Protection of Ukraine

Tsentralna str., 60, Dmytrivka vil., Ukraine, 08112

**Main Department of the State Emergency Service of Ukraine in Poltava Region

Reshetylvivska str., 26/1, Poltava, Ukraine, 36007

***Main Department of the State Emergency Service of Ukraine in Lviv Region

Pidvalna str., 6, Lviv, Ukraine, 79008

This work investigates the process of early fire detection by anomalous sources of thermal radiation registered by satellite remote sensing systems of the Earth. The subject of this study is anomalous sources of thermal radiation and their affiliation to an actual fire. The need for research is due to the fact that not every thermal anomaly recorded by a satellite corresponds to a real fire. Data on thermal anomalies were combined with the results of their ground verification in the Poltava oblast (Ukraine) in the period from 09/22/2020 to 04/23/2025. As a result of the combination, carried out on the basis of spatial and temporal proximity of points, 5096 clusters were established. At the same time, the share of clusters with confirmed fires was 13.0%. Each cluster was characterized by a set of 31 features, including spatial and temporal dimensions of the cluster; type of land cover; thermal radiation indicators, and meteorological indicators. A model has been built that determines the probability of a fire in a cluster for a given set of features. Using the method of recursive feature exclusion and a genetic algorithm, a set of 16 features was assembled, optimal according to the criterion of maximization of the PR AUC (Precision-Recall Area Under the Curve) metric.

A binary classifier has been built to identify fires among thermal anomalies. A feature of the model is its training on the results from ground verification of such anomalies. The model captures 85% of actual fires at a false detection rate of 54%.

In practice, this opens up opportunities for early detection of fire sites by ranking the detected sources of anomalous thermal radiation according to the criterion of fire probability. This makes it possible to improve the efficiency of using available forces and resources

Keywords: natural fires, satellite sensing, thermal anomalies, ground verification, genetic algorithm

Received 03.08.2026

Received in revised form 07.09.2026

Accepted 16.10.2026

Published 28.10.2026

How to Cite: Basmanov, O., Ohurtsov, S., Ponomarenko, R., Khyzhniak, A., Lysenko, O., Ryzhkov, Y., Chalyy, D., Chala, I., Hryshchenko, A. (2026). Building a model of early fire detection using machine learning methods and satellite imagery. *Eastern-European Journal of Enterprise Technologies*, 5 (10 (143)), 37–48. <https://doi.org/10.15587/1729-4061.2026.376307>

1. Introduction

Wildfires are among the most dangerous natural phenomena, reaching particularly large scales during dry periods [1]. One of the most negative consequences of forest

fires is the reduction of vegetation cover and the death of animals. The spread of fires to residential and other facilities leads to significant material damage and creates a direct threat to human life and health as well [2]. In addition, such large-scale fires are a significant source of

emissions of harmful substances into the atmosphere and their further spread [3].

As the area of a fire increases, the complexity of its localization and elimination increases, so the most effective approach is to detect and eliminate the fire at the initial stage of its evolution. One of the promising methods for early detection of fires of various types is spectral analysis of radiation [4]. Regarding natural fires, this approach is used in satellite monitoring. NASA's Fire Information for Resource Management System (FIRMS) uses data from the Moderate Resolution Imaging Spectroradiometer (MODIS) and Visible Infrared Imaging Radiometer Suite (VIIRS) satellite instruments. These data allow the detection of active fire sites and thermal anomalies and are updated every few hours [5]. Each record in the FIRMS database corresponds to a geographic point at which satellites recorded thermal radiation, which could potentially indicate the presence of a fire site [6]. However, the fact of registering a thermal anomaly does not guarantee that it corresponds to a fire site, since other sources of thermal radiation may also cause it.

The risk of forest fires is determined by a combination of natural and anthropogenic factors. The main natural factors include relative humidity, the condition and density of vegetation cover, which can be characterized by the normalized difference vegetation index (NDVI) [7], as well as the amount of precipitation. An important role is also played by the anthropogenic factor, in particular human activity in areas prone to fires.

Global trends indicate a possible increase in the area covered by forest fires by up to 30% due to anthropogenic climate change. The most pronounced trends are observed in temperate and boreal natural zones [8]. This emphasizes the need to devise regional approaches to assessing fire danger and early detection of fires.

Despite the availability of satellite data, their practical use remains difficult due to the uncertainty of which of the detected thermal anomalies correspond to actual fires. At the same time, ground-based verification of all detected anomalies is faced with a shortage of available resources.

Therefore, it is a relevant task to construct models for predicting the risk of fires in places of anomalous thermal radiation recorded using satellite monitoring.

2. Literature review and problem statement

2693 MODIS points were studied in [9], corresponding to 9 different fires that occurred in the state of Victoria (Australia) during the first three months of 2009. Remote sensing data were integrated with meteorological data. That made it possible to establish that the most significant factors affecting the probability of fire occurrence are soil moisture and temperature, air temperature and humidity, atmospheric pressure, and the vertical component of wind speed. The disadvantage of this approach is that conditions are considered only at those points where the fire occurred. In [10], the integration of remote sensing results with meteorological and other data was studied. It was found that the most significant factor is the type of land cover. This approach was applied to areas affected by fire. In this case, the issue of fire detection was left out of consideration. The results reported in [11] show that forest fires in southwestern China are most influenced by meteorological conditions, in particular the number

of clear sunny days, relative humidity (seasonal and daily), seasonal precipitation, and daily surface temperature. This approach characterizes the fire hazard of territories but does not solve the problem of operational classification of individual satellite thermal anomalies as fire sites or other sources of thermal radiation. The value of temporal and spatial parameters for forest fire forecasting was investigated in [12]. Similarly, the task of verification of an individual thermal anomaly is not considered.

The McArthur Forest Fire Danger Index (Australia), the National Fire Danger Rating System (USA) [13], and the Canadian Forest Fire Weather Index (Canada) are widely used to assess the fire hazard level of forest areas. The Canadian index is calibrated to describe the behavior of fires in Banks pine (*Pinus banksiana*) stands, which are typical for Canadian forests [14]. This complicates the application of such indices outside the regions for which they were developed and necessitates devising new indices adapted to specific conditions.

The processes of natural fire evolution are characterized by complexity and interaction of various factors, in particular the type of combustible material, wind speed, temperature, humidity, and other parameters [15]. In [16], existing machine learning approaches to predicting and detecting forest fires are systematized, but the task of classifying satellite thermal anomalies as "fire" or "non-fire" is left out. In [17], machine learning algorithms, in particular the XGBoost (Extreme Gradient Boosting), Random Forest, and Multilayer Perceptron methods, were used to predict the daily rate of fire spread based on meteorological factors, terrain characteristics, and fuel stocks. However, the task of verifying the sources of anomalous thermal radiation was not considered here either. In [18], the MODIS Global Fire Atlas was used as a source of data on fire occurrence, and the Random Forest algorithm was used to determine the factors contributing to the occurrence of forest fires. In this case, the sample is formed only from registered fires, and not from all detected thermal anomalies.

In [19], the hyperparameters of the XGBoost model were optimized to predict the rate of increase in the area of a forest fire. However, the task of verifying the detected sources of anomalous thermal radiation was not considered. In [20], machine learning algorithms were used to construct a fire hazard index. This approach allows one to assess the favorableness of weather conditions for the occurrence of a fire, rather than to determine whether a satellite-recorded thermal anomaly is a real fire. In [21], maps of the area's susceptibility to forest fires were constructed based on known ignition points. However, this approach does not solve the task of classifying individual satellite thermal anomalies according to their correspondence to actual fires. In [22], it was noted that taking into account the characteristics of the terrain and using machine learning models to predict the likely area of fire spread provided results that exceeded the quality of the results obtained using the FARSITE fire modeling tool.

In [23], the effect of different combinations of features on the efficiency of the XGBoost model in predicting forest fire intensity was investigated. In [24], a genetic algorithm was used to determine the optimal combination of features. However, those studies did not consider the task of initial verification of satellite thermal anomalies. In [25], logistic regression and the XGBoost method were used to classify satellite pixels to determine the boundaries of burned forest areas. For this

purpose, MODIS observation results, data on active fires, and other spatial features were used. The reference sample was formed based on data on burned forest areas. However, the task of early fire recognition among satellite thermal anomalies was left out of consideration.

Thus, existing approaches based on satellite data largely focus on the characteristics of already detected fires and their severity assessment. At the same time, the task of automated classification of satellite thermal anomalies corresponding to actual fires from other thermal anomalies remains a separate issue.

The above allows us to argue that it is advisable to conduct a study aimed at building an early fire detection model based on data on sources of anomalous thermal radiation and the results of their verification regarding the presence of a fire.

3. The study materials and methods

The aim of our work is to construct an early fire detection model based on satellite monitoring data on sources of anomalous thermal radiation and the results of their verification for the presence of a fire. In practice, this opens up opportunities for early detection of fire sites by ranking the detected sources of anomalous thermal radiation according to the criterion of fire probability. This, in turn, makes it possible to improve the efficiency of using available forces and resources.

To achieve this aim, the following objectives were accomplished:

- to integrate MODIS and VIIRS data on anomalous thermal radiation with satellite data on land cover types, meteorological indicators, and State Emergency Service data on actually recorded fires;
- to estimate the conditional probability of fire presence based on the characteristics of thermal anomaly centers based on the XGBoost algorithm;
- to determine the most informative parameters and evaluate the effectiveness of the model using forecasting quality metrics.

4. The study materials and methods

The object of our study is the process of early detection of fires by anomalous sources of thermal radiation registered by satellite systems of remote sensing of the Earth. The subject of the study is anomalous sources of thermal radiation and their affiliation to an actual fire. The principal hypothesis assumes that the combination of satellite, meteorological and spatial-temporal data makes it possible to determine the probability of a fire by anomalous thermal sources. Basic assumptions are the possibility of using the identified patterns for similar conditions and the reliability of ground verification by the State Emergency Service of anomalous sources of thermal radiation. For simplification, at the sites of anomalous thermal radiation, not actual meteorological conditions were used, but data provided by weather forecast services.

The study was conducted in the territory of Poltava oblast, Ukraine. Poltava oblast is located in the central and north-eastern parts of Ukraine and almost completely covers the territory of the forest-steppe natural zone. Its total area is approximately 28,750 km². The main natural plant communities are forests, meadows, swamps, and steppes.

In the period from 09/22/2020 to 04/23/2025, all FIRMS data on anomalous sources of thermal radiation in the Poltava oblast were sent to the information system at the Main Directorate of the State Emergency Service of Ukraine in the Poltava oblast. After that, they were checked by directly sending units to the location of each potential source of fire (Fig. 1). The exception was the period from February 24, 2022, to April 23, 2022, when no observations were made. As a result, a database was assembled, which contained the following fields:

- latitude and longitude of anomalous source of thermal radiation;
- the time of receiving information about a potential fire;
- indicator of the presence (1) or absence (0) of an actual fire.

The study used archival FIRMS data [6] for the period from 09/22/2020 to 02/23/2022 and from 04/24/2022 to 04/23/2025. Only those points belonging to the Poltava oblast were selected from the FIRMS data. For this purpose, its administrative border was approximated by a closed broken line [26]. After that, the location of the point inside the formed polygon was checked. Further analysis was carried out exclusively for points located within the Poltava oblast.

We took into account the following characteristics:

- latitude and longitude of the source of anomalous thermal radiation;
- date and time of observation;
- temperature in the infrared spectrum (K);
- background temperature (K);
- thermal power of the fire (MW).

Information on land cover types was obtained from the ESRI platform [27], which provides data from remote sensing of the Earth by Sentinel-2 satellites. Based on these data, the following land cover types are distinguished:

- water bodies;
- trees;
- flooded vegetation;
- agricultural land;
- built-up areas;
- open soil;
- areas with permanent snow cover;
- areas with permanent cloud cover;
- pastures.

The following meteorological factors were investigated that could potentially affect fire danger:

- air temperature;
- air humidity;
- precipitation;
- wind speed;
- cloud cover.

The online weather forecasting service Weather Forecast [28] and the Google Earth Engine platform (USA) [29] were used as sources of meteorological data. The Google Earth Engine platform was also used to acquire the NDVI and NDMI vegetation indices.

All stages of model construction were implemented programmatically using the Python programming language (The Netherlands) (version 3.13) using a set of specialized libraries. The XGBoost model was chosen as the basic classification algorithm. To reduce the dimensionality of the feature space and select the most informative ones, the Recursive Feature Elimination (RFE) method was used. The optimal set of features was determined using a genetic algorithm.

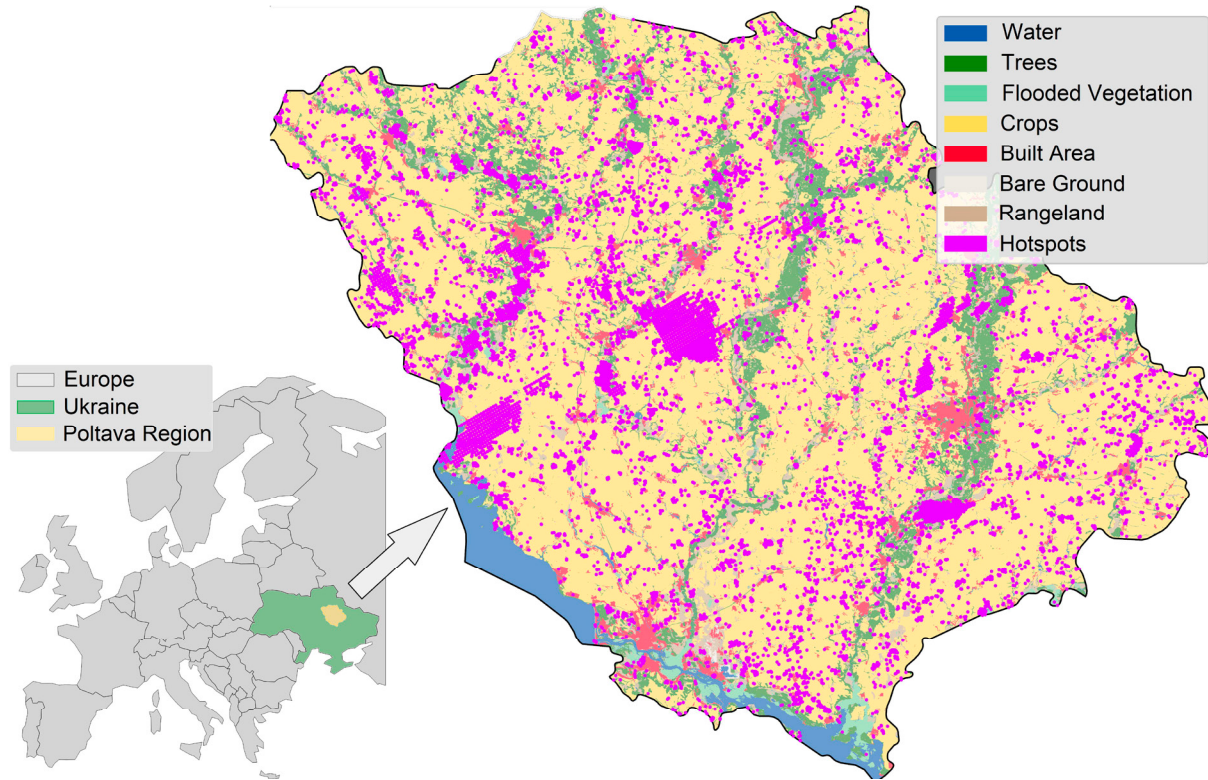


Fig. 1. Study area: Poltava oblast, Ukraine

5. Results of building a model for early detection of fire sites using satellite data

5.1. Integration of satellite and meteorological data with data on the type of land cover

Taking into account the spatial resolution of satellite data (Table 1), the accuracy in determining the coordinates of a potential fire site is ± 1.0 km or ± 0.375 km, depending on the satellite.

Table 1

Satellites used for fire monitoring

Satellite	Detector	Resolution, km	Launch year
Terra	MODIS	1.0	1999
Aqua	MODIS	1.0	2002
Suomi-NPP	VIIRS	0.375	2011
NOAA-20	VIIRS	0.375	2017
NOAA-21	VIIRS	0.375	2022

The criterion for combining two given points from the FIRMS database into one cluster was their spatial and temporal proximity:

$$L(\phi_1, \lambda_1, \phi_2, \lambda_2) \leq L_{\max}; \quad (1)$$

$$|t_1 - t_2| \leq t_{\max}, \quad (2)$$

where λ_1, λ_2 are the longitudes of the points; ϕ_1, ϕ_2 are the latitudes of the points; t_1, t_2 are the time of fire detection. The distance between points on the Earth's surface was determined from the following formula

$$L = 2R \times \arcsin \left(\sqrt{\sin^2 \left(\frac{\phi_1 - \phi_2}{2} \right) + \cos \phi_1 \cos \phi_2 \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right)} \right), \quad (3)$$

where R is the radius of the Earth.

Taking into account the error in determining the spatial coordinates of the thermal radiation source, the maximum distance $L_{\max} = 0.5$ km was chosen for the clustering procedure. A certain point can be included in satellite images approximately 16 times per day, which corresponds to an average interval between images of about 1.5 hours. However, in practice, satellite images are distributed unevenly in time. Therefore, the maximum time interval $t_{\max} = 3$ hours was chosen.

The State Emergency Service database contains the date and time of entering information into its own information system but does not record the corresponding time in the FIRMS system. Analysis of this data set revealed that the interval between consecutive entries in the information system could be up to 12 hours. Therefore, when adding this data to clusters, the spatial criterion (1) was used, and the following inequality was used as the time criterion

$$-t_{\max_1} \leq t_2 - t_1 \leq t_{\max_2}, \quad (4)$$

where t_1 – time of registration of a probable fire in the FIRMS system; t_2 – time of entering information about a probable fire into the State Emergency Service information system; $t_{\max_1} = 3$ h; $t_{\max_2} = 12$ h.

The input data consisted of:

- 15138 records from the FIRMS database;
- 16356 records from the State Emergency Service database.

As a result of clustering, the following results were obtained:

- 1635 records from the FIRMS database (11%) did not have a corresponding record in the State Emergency Service database;
- 2993 records from the State Emergency Service database (18%) did not have a corresponding record in the FIRMS database;
- 13503 records from the FIRMS database and 13363 records from the State Emergency Service database were combined into 5096 clusters, each of which contained at least one record from the FIRMS database and one record from the State Emergency Service database (Fig. 2).

Among the FIRMS database records that are not linked to any record in the State Emergency Service of Ukraine database, most (1022 records) are individual records that are not clustered with other records.

Among the State Emergency Service of Ukraine database records that are not linked to any record in the FIRMS database, most (2041 records) are also individual records that are not clustered with other records. Among these records, fire was confirmed in 106 cases, which is 3.5%. At the same time, the share of confirmed fires among all State Emergency Service of Ukraine data is 22%.

Thus, among the State Emergency Service of Ukraine database records that are not linked to any FIRMS record, the share of confirmed fires is approximately six times lower than the average value for the entire database. In addition,

satellite observation data (in particular, source and background temperature or thermal radiation power) are missing for these records. Therefore, they were excluded from further analysis.

Similarly, FIRMS records that were not associated with any record in the State Emergency Service database were also excluded from further analysis. This was done because they lacked ground verification results that would allow us to determine whether the detected anomalous thermal radiation corresponded to the occurrence of a fire.

Therefore, further analysis was conducted using 13,503 FIRMS records and 13,363 State Emergency Service database records. These records were combined into 5,096 clusters, each containing at least one FIRMS record and one State Emergency Service database record. A cluster was considered to contain a fire if it included at least one confirmed fire.

$$y = \begin{cases} 1, k > 0; \\ 0, k = 0, \end{cases} \quad (5)$$

where k is the number of confirmed fire sites.

While burning was confirmed for 26.2% of all points in the State Emergency Service of Ukraine database included in the clusters, the share of clusters with confirmed burning was 13.0%. This discrepancy was due to a large fire that lasted from September 15 to 17, 2024 and covered 721 records with confirmed burning.

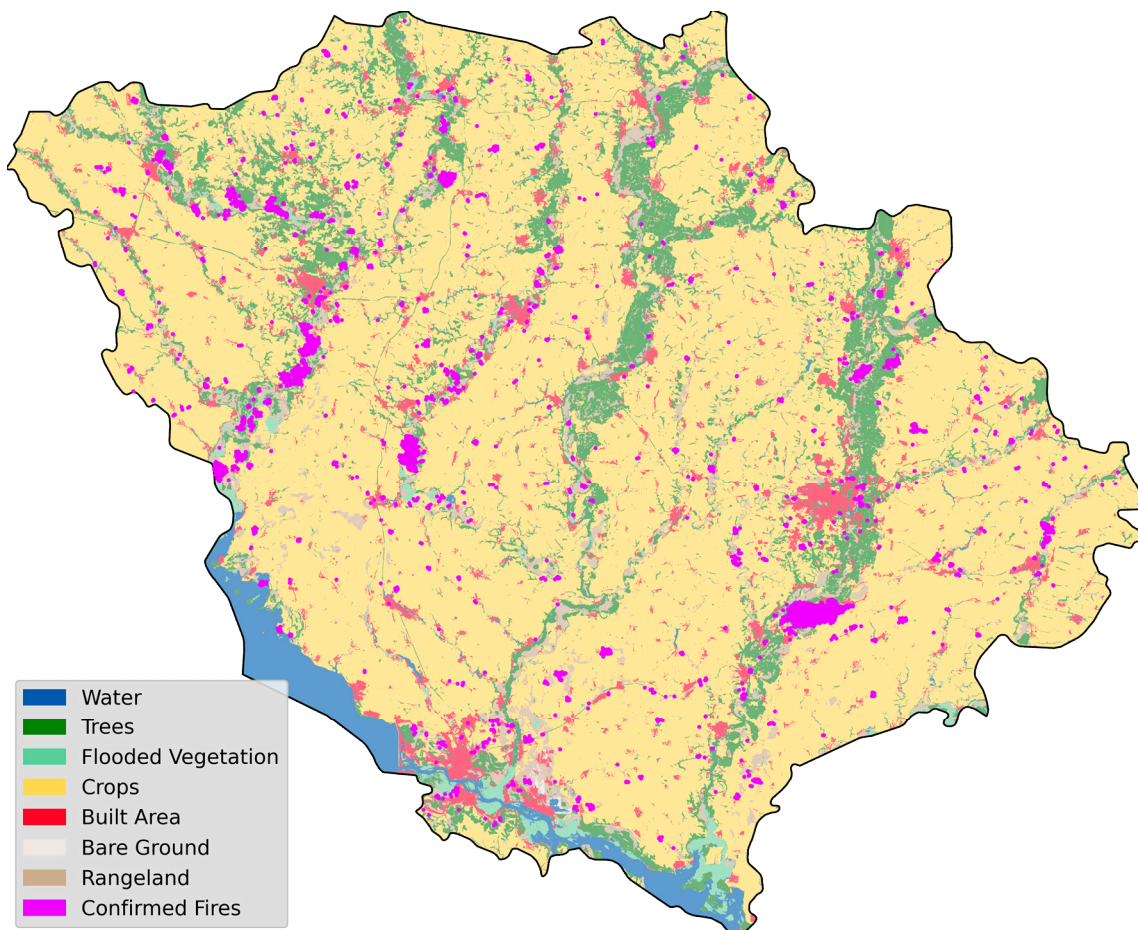


Fig. 2. Confirmed fires

5. 2. Estimation of the conditional probability of fire presence based on the characteristics of thermal anomaly sites based on the gradient boosting algorithm

To build the model, a set of features was selected that characterize the location of the cluster, its size, and the thermal radiation of the source (FIRMS data). Weather conditions obtained from the Google Earth Engine (GEE) and Weather Forecast (WF) services, the type of surface cover (ESRI), as well as the NDVI and NDMI indices (GEE), were added to them.

To take into account the cluster size, the following features were used: the number of FIRMS data points within the cluster, the spatial and temporal dimensions of the cluster. The spatial size of the cluster was defined as the area of the smallest rectangle that covers all FIRMS points

$$S = \left(\max_{i=1..m} x_i - \min_{i=1..m} x_i \right) \times \left(\max_{i=1..m} y_i - \min_{i=1..m} y_i \right), \quad (6)$$

where (x_i, y_i) are the coordinates of the FIRMS points, expressed in kilometers; m is the number of points in the cluster.

The temporal size of a cluster is defined as the time interval (in hours) between the moments of registration of the first and last points in the FIRMS database

$$t = \max_{i=1..m} t_i - \min_{i=1..m} t_i, \quad (7)$$

where t_i is the point registration time in the FIRMS database.

Season and time of day also affect the likelihood of a fire. These features can be represented on a linear scale: the days of the year are from 1 to 365, and the hours of the day are from 0 to 24. However, they are cyclical in nature. For example, the difference between the 365th day and the 1st day of the year is actually one day, not 364 days as the linear representation suggests. Therefore, to take into account the cyclic nature of these variables, the model used features obtained by sine transformation:

$$n_{ds} = \sin \frac{\pi n_d}{365}, \quad (8)$$

$$t_s = \sin \frac{\pi t}{24}, \quad (9)$$

where n_d is the day number of the year: from 1 to 365 for a non-leap year and from 1 to 366 for a leap year; n_{ds} is its sine transformation; t is the time of day; t_s is its sine transformation.

The total number of features pre-selected for building the model was 31 (Table 2). Pairwise correlation analysis revealed a high level of correlation between some features (more than 0.9). Therefore, the maximum wind speed in the cluster and the spatial size of the cluster were excluded from further consideration. As a result, the number of features used in the study was reduced to $N = 29$.

The type of cover was represented by separate features, each of which corresponded to a certain type of cover and acquired the value 0 or 1. The angle of inclination and its orientation were not included in the set of features, since the studied area is located on a flat terrain.

To predict the occurrence of fire, a probabilistic binary classification model ("fire" or "no fire") was used. This model determines the conditional probability of belonging to the positive class "fire" for a given set of features

$$p = P(y=1|X) = f(x), \quad (10)$$

where the target variable y is a binary indicator function (5); X is a feature vector characterizing cluster properties; f is a learning function that approximates the dependence of the probability of fire occurrence on the feature vector X .

Table 2

Types of features and their sources

No.	Feature	Source
1	Latitude of the cluster center	FIRMS
2	Longitude of the cluster center	FIRMS
3	Number of points in the cluster	FIRMS
4	Time size of the cluster	FIRMS
5	Maximum temperature in the cluster	FIRMS
6	The maximum relative temperature in the cluster	FIRMS
7	Maximum thermal radiation in the cluster	FIRMS
8	Average background temperature	FIRMS
9	Average daily air temperature	GEE
10	Maximum air temperature in the cluster	WF
11	Maximum air humidity in the cluster	WF
12	Average daily difference between air temperature and dew point	GEE
13	Maximum precipitation in the cluster	WF
14	Average precipitation over the last 5 days	GEE
15	Cloud cover	WF
16	Maximum atmospheric pressure	WF
17	Total solar radiation for the last 5 days	GEE
18	Average daily wind speed	GEE
19	Vegetation index	GEE
20	Average humidity index for the last 5 days	GEE
21	Cover type – water	ESRI
22	Cover type – trees	ESRI
23	Cover type – flooded vegetation	ESRI
24	Cover type – agricultural land	ESRI
25	Cover type – buildings	ESRI
26	Cover type – pasture	ESRI
27	Sine transform for day of year	(8)
28	Sine transform for time of day	(9)
29	Weekends	Calendar
30	Maximum wind speed in cluster	WF
31	Spatial size of cluster	FIRMS

An ensemble method of gradient boosting based on decision trees (XGBoost) was used to construct the function f . After training, model (10) returns the value p , $0 \leq p \leq 1$, which can be interpreted as the probability of fire occurrence in the corresponding cluster. Entering a certain threshold value m_{th} made it possible to obtain a binary classifier that allows the result to be interpreted as a fire or its absence:

$$g = \begin{cases} 1, p \geq m_{th} \Rightarrow \text{fire;} \\ 0, p < m_{th} \Rightarrow \text{no fire.} \end{cases} \quad (11)$$

The effectiveness of the model was evaluated by the following metrics:

– precision

$$p_p(m_{th}) = \frac{n_{tp}}{n_{tp} + n_{fp}}; \quad (12)$$

– recall

$$p_r(m_{th}) = \frac{n_{tp}}{n_{tp} + n_{fn}}; \quad (13)$$

– false positive rate

$$p_{fpr}(m_{th}) = \frac{n_{fp}}{n_{fp} + n_{tn}}, \quad (14)$$

where n_{tp} is the number of true positive realizations (the model predicted a fire and there was a fire); n_{tn} is the number of true negative realizations (the model predicted the absence of a fire and there was no fire); n_{fp} is the number of false positive realizations (the model predicted a fire, but there was no fire); n_{fn} is the number of false negative realizations (the model predicted the absence of a fire, but there was a fire).

Formulas (12) and (13) give the dependence of precision on completeness $p_p(p_r)$ – the Precision-Recall (PR) metric and the integral metric

$$PR AUC = \int_0^1 p_p(p_r) dp_r. \quad (15)$$

Similarly, formulas (13) and (14) give the dependence of completeness on the fraction of false positives $p_r(p_{fpr})$ – the receiver operating characteristic (ROC) and the integral metric

$$ROC AUC = \int_0^1 p_r(p_{fpr}) dp_{fpr}. \quad (16)$$

To assess the model's ability to learn and control over-training, the original dataset was divided into two disjoint subsets: training and testing. The distribution was performed randomly while maintaining the initial ratio between positive and negative classes. The ratio of the training and testing sample sizes was 80% to 20%.

To assess the model's performance on the testing sample, the integral metrics PR AUC and ROC AUC were calculated. However, only the PR AUC metric was used as an optimization criterion when selecting model parameters. Training the model with an initial number of features ($N = 29$) yielded the following model performance estimates: PR AUC = 0.639; ROC AUC = 0.898.

5.3. Determining the most informative parameters of the model and assessing its effectiveness

The method of recursive feature elimination (RFE) was used to pre-select informative features and reduce the dimensionality of the feature space. At each iteration of the RFE procedure, one feature was eliminated from the current set of features – the one that had the lowest relative importance according to the results of training the XGBoost model. After removing the feature, the model was retrained on the remaining features, after which the PR AUC and ROC AUC metrics were calculated on the test sample. Figure 3 shows the dependence of the PR AUC and ROC AUC values on the number of features used in the model.

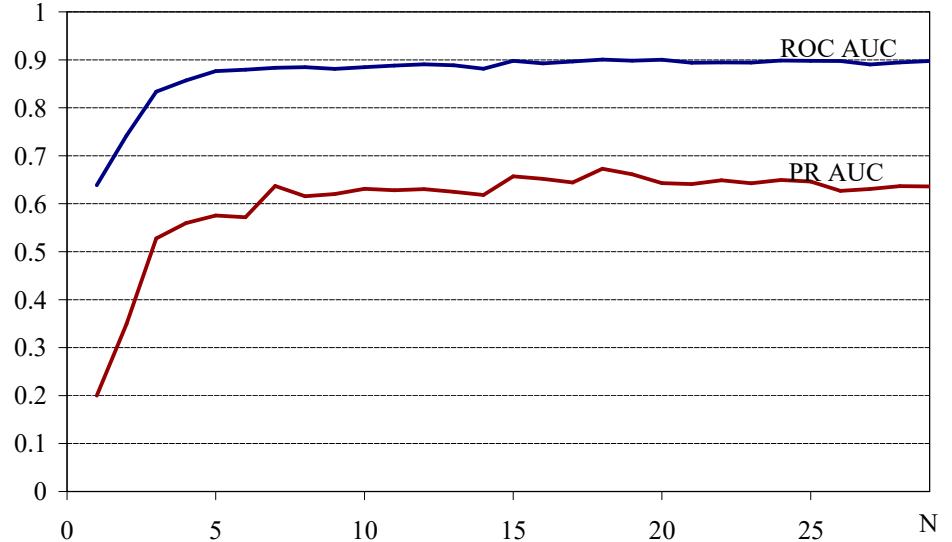


Fig. 3. Dependence of the metrics Precision-Recall Area Under the Curve and Receiver Operating Characteristic – Area Under the Curve on the number of model features

Analysis of the dependences shown in Fig. 3 reveals that the PR AUC metric acquires the highest values for the number of model features from 15 to 19. The maximum value of 0.6729 is achieved at $N = 18$. Further increase in the number of features did not lead to an increase in the efficiency of the model.

To select the optimal set of features that provides the maximum PR AUC value, a genetic algorithm from the DEAP library (Python) was used. The number of features in the final solution was allowed to vary from 12 to 25. The evolutionary process included selection, crossover, and mutation operations aimed at gradually increasing the overall predictive ability of the model.

As an initial population, feature sets previously selected using the RFE method were used. The fitness of each individual was assessed based on the PR AUC values obtained during training the model (10) with the corresponding set of features. To reduce the influence of random factors, stratified 3-fold cross-validation was used, the essence of which was to divide the original data set into 3 identical groups. Within each group, the model was trained on the training sample. After that, the PR AUC value was calculated on the test sample. The final value of the fitness function was determined as the average PR AUC value for all test samples.

Based on the results of applying the genetic algorithm, 16 features were selected (Table 3).

Fig. 4 illustrates the influence of the features listed in Table 3 on the model predictions using the SHAP (Shapley Additive Explanations) methodology [30].

Each point reflects the contribution of a separate feature to the prediction of the conditional probability of fire occurrence. The abscissa corresponds to the SHAP value: positive values increase the predicted probability of fire occurrence, while negative values decrease it. The color coding of the points corresponds to the empirical value of the feature: red indicates high values, blue indicates low values.

For example, high values of the feature “crops” are associated with negative SHAP values. Thus, in the case of detecting an anomalous point of thermal radiation on agricultural land (crops = 1), the model determines a lower predicted probability of fire occurrence; on the other hand,

for crops = 0, the predicted probability of fire increases. A comparison of the images in Fig. 1, 2 confirms that for most of the anomalous radiation sources detected on agricultural land, the presence of fire was not confirmed.

The performance characteristics of the constructed model are shown in Fig. 5. In this case, the values of the metrics were as follows: PR AUC = 0.68; ROC AUC = 0.90.

To obtain the binary classifier (11), optimization was performed using the threshold value m_{th} . The threshold value was determined by the criterion

$$p_p(m_{th}) \rightarrow \max, \quad (17)$$

under condition

$$p_r(m_{th}) \geq 0.85. \quad (18)$$

This approach provides high sensitivity of the classifier while maintaining an acceptable level of false positives. The threshold selection procedure was based on constructing the Precision-Recall curve for the validation subsets (Fig. 5) and further determining the point corresponding to the maximum p_p value among all threshold values m_{th} for which the condition $p_r(m_{th}) \geq 0.85$ is satisfied.

As a result, the threshold value $m_{th} = 0.089$ was obtained. The parameters characterizing the efficiency of the classifier are given in Table 4.

In Table 4, “precision” means the proportion of cases correctly assigned to a given class among all cases assigned to a given class. “Recall” is the proportion of cases correctly assigned to a given class among all cases that should have been assigned to a given class.

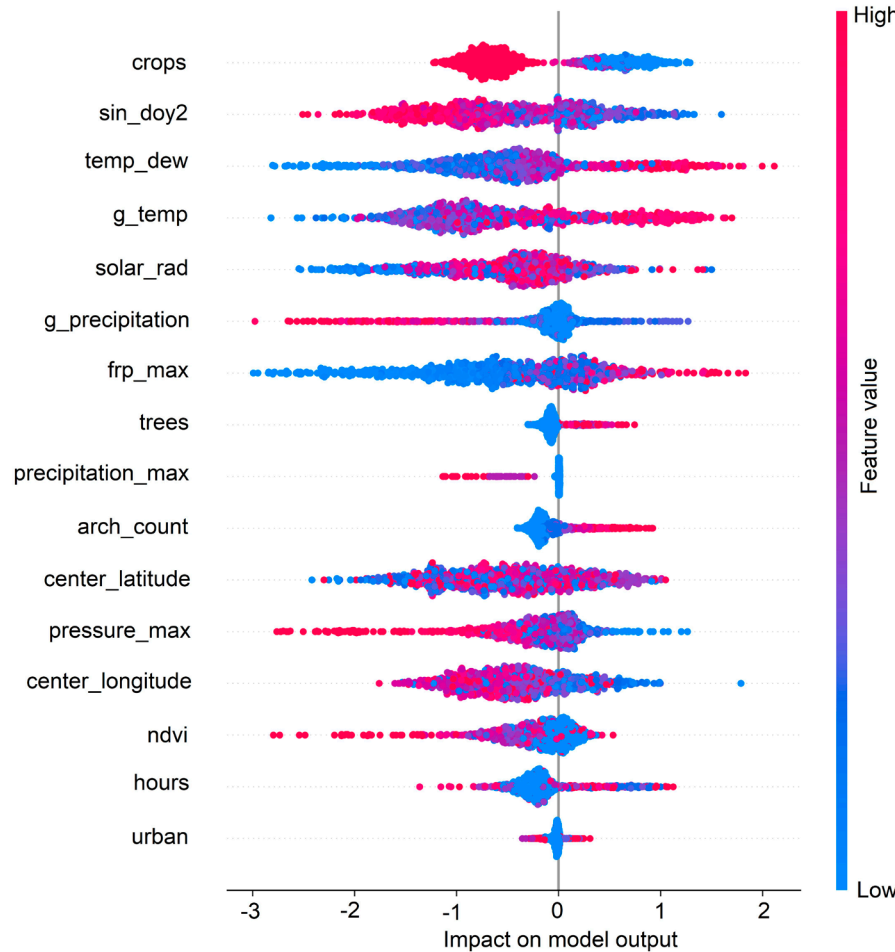


Fig. 4. Distribution diagram (beeswarm) of Shapley Additive Explanations values for key features in the gradient boosting model

Table 3

The set of features that are optimal according to the criterion of maximizing the PR AUC metric

Feature	Abbreviation	Significance
Agricultural land	crops	0.174363
Sine transform of day of year number	sin_doy	0.079859
Average daily difference between air temperature and dew point	temp_dew	0.078908
Average daily air temperature	g_temp	0.069388
Total solar radiation for the last 5 days	solar_rad	0.064594
Average precipitation for the last 5 days	g_precipitation	0.063485
Maximum thermal radiation power of the cluster	frp_max	0.058805
Trees	trees	0.049542
Maximum precipitation in the cluster	precipitation_max	0.049341
Number of points in the cluster	arch_count	0.048910
Latitude of the cluster center	center_latitude	0.048128
Maximum atmospheric pressure in the cluster	pressure_max	0.047314
Longitude of the cluster center	center_longitude	0.046817
Vegetation index	ndvi	0.043566
Temporal size of the cluster	hours	0.043552
Buildings	urban	0.033427

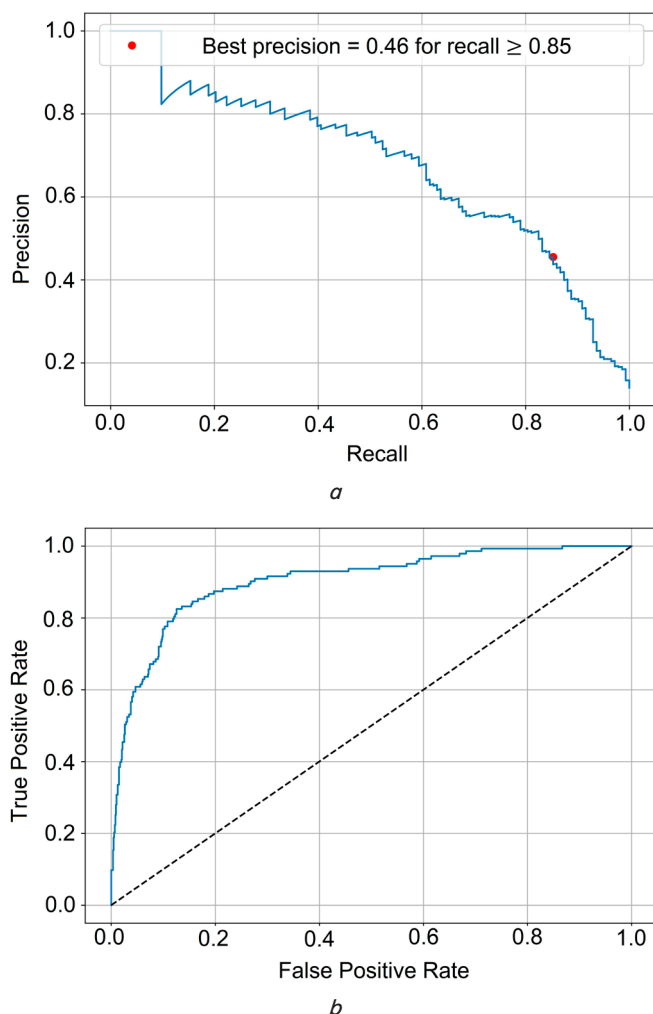


Fig. 5. Performance characteristics of the constructed model: *a* – Precision-Recall curve; *b* – receiver performance characteristics

Table 4

Classifier efficiency for threshold value $m_{th} = 0.089$

Class	Precision	Recall
Fire	0.4552	0.8531
No fire	0.9721	0.8335

6. Discussion of results based on the construction of a model for early detection of fire sites

The accuracy in satellite determination of the source of thermal radiation is ± 1.0 km or ± 0.375 km (Table 1). This means that the same source of thermal radiation on the earth's surface may have different coordinates on different satellite images. In addition, in the case of forest and steppe fires, the source of thermal radiation is not limited to one point but covers a certain area. Thus, one fire may correspond to a group of spatially close points. In view of this, the obtained points were clustered using the criteria of spatial (1) and temporal (2) proximity. This approach, on the one hand, makes it possible to assess the spatial and temporal scale of a potential fire, and on the other hand, to compensate for errors in determining the coordinates of the source of thermal radiation.

To add points from the State Emergency Service database to these clusters, the spatial proximity criterion (1) was used, and the temporal proximity criterion was replaced by (4). The need for such a replacement was due to the fact that the current time was entered into the State Emergency Service database, and not the time of registration of thermal radiation by a satellite. This allowed us to select 5096 clusters that contained at least one point from the FIRMS and State Emergency Service databases. The presence of a confirmed fire in a cluster was determined by function (5).

Meteorological characteristics and the type of land cover were added to the spatial (6) and temporal (7) dimensions of the cluster. The season was also taken into account by formula (8) and the time of day (9). Among the 31 features obtained (Table 2), pairs of features were found that correlate with each other with a correlation coefficient of more than 0.9. This allowed us to exclude one feature from each pair and reduce the previous number of features to 29. The obtained features can be used to predict the presence of a fire in a cluster. The argument of the function for predicting a fire in a cluster (10) is the value of these features, and the result is the conditional probability of a fire. Introducing a threshold value for the conditional probability of a fire allowed us to divide the range of values of function (10) into two subranges and obtain a binary classifier (11). The classifier (11) returns 1 if the model predicts a fire in the cluster, and 0 otherwise.

The choice of a set of features, which is the argument of function (10), affects the values of model efficiency indicators (12)–(16). In this case, an increase in the values of some efficiency indicators is accompanied by a decrease in the values of others. In addition, the indicators (12)–(14) also depend on the choice of the threshold value for the classifier (11). The initial data are characterized by a significant imbalance of classes: the positive class “fire” occurs much less often (13%) than the negative class “no fire” (87%). At the same time, the cost of errors is not the same: a false negative result (missing a fire) is more dangerous than a false positive (false call to a fire). Therefore, the integral PR AUC metric (15) was used as the model efficiency metric. This approach is consistent with the results from previous studies, which indicated that PR AUC is a more informative metric than ROC AUC in classification problems with unbalanced classes [31].

Not all of the 31 pre-selected features are equally important for the fire prediction model (10). Moreover, some features may introduce additional noise, reducing the efficiency of the model. The use of the Recursive Feature Elimination (RFE) algorithm to sequentially remove the least significant features allows us to analyze the efficiency of the model depending on the number of features used (Fig. 3). In this case, the maximum value of the PR AUC metric occurs for the number of features $N = 18$. At the same time, there is a fairly wide plateau in both directions, on which the PR AUC value is close to the maximum.

The feature sets pre-selected using the RFE algorithm were used as part of the initial population for the genetic algorithm. Unlike RFE, which sequentially removes one feature for each iteration, the genetic algorithm retains and combines various sets of features. This contributes to the emergence of locally optimal configurations [32]. As a result of the application of the genetic algorithm, 16 features were selected for inclusion in the model (10) (Table 3). Their analysis reveals that the most

important factors determining the effectiveness of fire forecasting are the type of land cover, season, and air humidity.

The Beeswarm plot (Fig. 4) provides a visual interpretation of the contribution of each feature to the fire prediction model (10). In particular, land cover type is a key factor: high values of the “crops” feature (red color) are mainly associated with negative SHAP values, which leads to a decrease in the predicted probability of fire occurrence. This is consistent with empirical evidence that most thermal anomalies in agricultural lands (Fig. 1, 2) correspond to events not related to burning. In contrast, the “trees” feature is characterized by positive SHAP values, which increase the predicted fire risk in tree-covered areas. This pattern is consistent with fire regimes in Eurasia, where seasonal decreases in humidity contribute to increased availability of combustible material [33].

The seasonal cyclicality, represented by the sine transform (8), is characterized by a concentration of positive SHAP values in late summer (days 200–250). These days are characterized by increased solar exposure and decreased precipitation. Air humidity and the difference between air temperature and dew point temperature also have a significant impact: low humidity values increase the predicted probability of fire occurrence, while high humidity decreases this probability.

The results obtained using SHAP confirm the effectiveness of the combined use of RFE and a genetic algorithm for feature selection, revealing nonlinear dependences that are not taken into account by linear indices such as FWI [14]. Our results are consistent with studies [34], which established the high importance of meteorological factors. The fact that the NDMI index was not included in the constructed model is due to the fact that it is derived from the type of land cover, season, precipitation, and humidity. All these factors are already represented in the model (Table 3). Unlike the models built for Australia, which focus on soil moisture [9], the proposed approach relies mainly on air humidity, land cover type, and season, which is consistent with the patterns characteristic of Eurasia [33].

The next step was to determine the probability threshold for the classifier (11). Unlike [35], in which the criterion of equality of probabilities of false detection and missed fire was used, the threshold was determined by maximizing the accuracy (17) under the condition of a given minimum level of completeness (18). This approach is justified by the practical need to minimize the probability of type II errors, since in the context of fire monitoring, missing a fire source could lead to a delay in response and increased fire damage. The constructed model (Table 4) is capable of detecting 85% of actual fires, while the proportion of false detections is about 54%. This model configuration focuses on a compromise between sensitivity to rare events and the ability to avoid false alarms.

A feature of the model is its training based on the results from ground verification of recorded thermal anomalies with additional consideration of the type of land cover and meteorological conditions.

Our results could be used to manage the risk of natural fires and devise resource allocation strategies. Preliminary calculation of conditional probabilities of fire occurrence (10) in the nodes of a given grid allows one to form a fire hazard index for a specific region. This index takes into account the unique characteristics of the relevant territory. The method of recursive feature exclusion and genetic algorithm could be used in the design of self-adjusting fire detectors [36].

A limitation of the proposed approach to building an early fire detection model is the availability of a database of actual fires that occurred in a given area over a certain period of time.

The disadvantage of our study is that it was conducted under certain natural and climatic conditions, which makes it impossible to directly apply the results under other conditions.

Prospects for further research are related to improving the model by including historical fire data or distances to infrastructure facilities instead of using raw geographic coordinates.

7. Conclusions

1. NASA FIRMS satellite data on thermal anomalies on the Earth’s surface have been integrated with verified data from the State Emergency Service of Ukraine on recorded fires that occurred in the Poltava oblast from September 22, 2020 to April 23, 2025. The integration is based on combining individual points into clusters based on the criteria of spatial and temporal proximity. Taking into account the resolution of satellite images, points with a distance between them of no more than 500 m were assigned to one cluster. Out of 15,138 records from the NASA FIRMS database and 16,356 records from the State Emergency Service database, 5,096 clusters were obtained. At the same time, the share of clusters with confirmed burning was 13.0%.

2. A set of 31 features was assembled, including spatial and temporal cluster size, land cover type, thermal radiation indicators according to NASA FIRMS, and meteorological indicators. To predict the presence of a fire in a cluster, the conditional probability of belonging to the “fire” class for a given set of features was estimated. Machine learning methods were used to build the estimate, namely, the gradient boosting method based on decision trees (XGBoost). The effectiveness of the forecast was assessed using the integral PR AUC metric. 2 features were excluded from the pre-formed set, for which the values of pairwise correlations with other features exceeded 0.9.

3. Using the method of recursive feature exclusion and genetic algorithm, a subset of 16 features was obtained that determine the conditional probability of a fire in the cluster. Despite the reduction in the number of features by almost 2 times, the efficiency of the model by the PR AUC metric increased by 5.3%. The optimal probability threshold for a binary classifier was determined. The constructed model is capable of detecting 85% of actual fires with a false detection rate of about 54%. This model configuration provides a compromise between sensitivity to rare events and minimization of the false alarm rate.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study and the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

Data will be provided upon reasonable request.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

Author contributions

Oleksii Basmanov: Conceptualization, Software, Writing – original draft; **Serhii Ohurtsov:** Supervision, Writing – re-

view & editing; **Roman Ponomarenko:** Methodology, Project administration; **Andrii Khyzhniak:** Validation, Investigation; **Oleksandr Lysenko:** Formal analysis, Visualization; **Yevhen Ryzhkov:** Data curation; **Dmytro Chalyy:** Writing – review & editing, Software; **Iryna Chala:** Writing – review & editing; **Anastasiia Hryshchenko:** Writing – review & editing.

All authors have read and agreed with the published version of the manuscript and are jointly responsible for the reliability, accuracy, and integrity of the data and research results reported in the paper.

References

- Halofsky, J. E., Peterson, D. L., Harvey, B. J. (2020). Changing wildfire, changing forests: the effects of climate change on fire regimes and vegetation in the Pacific Northwest, USA. *Fire Ecology*, 16 (1). <https://doi.org/10.1186/s42408-019-0062-8>
- Saleh, A., Zulkifley, M. A., Harun, H. H., Gaudreault, F., Davison, I., Spraggon, M. (2024). Forest fire surveillance systems: A review of deep learning methods. *Heliyon*, 10 (1), e23127. <https://doi.org/10.1016/j.heliyon.2023.e23127>
- Popov, O., Taraduda, D., Sobyna, V., Sokolov, D., Dement, M., Pomaza-Ponomarenko, A. (2020). Emergencies at Potentially Dangerous Objects Causing Atmosphere Pollution: Peculiarities of Chemically Hazardous Substances Migration. *Systems, Decision and Control in Energy I*, 151–163. https://doi.org/10.1007/978-3-030-48583-2_10
- Pospelov, B., Rybka, E., Samoilov, M., Morozov, I., Bezuhla, Y., Butenko, T. et al. (2022). Defining the features of amplitude and phase spectra of dangerous factors of gas medium during the ignition of materials in the premises. *Eastern-European Journal of Enterprise Technologies*, 2 (10 (116)), 57–65. <https://doi.org/10.15587/1729-4061.2022.254500>
- Plank, S., Ciancia, E., Genzano, N., Falconieri, A., Martinis, S., Taubenböck, H. et al. (2025). The evolution of the 2022–2024 eruption at Home Reef, Tonga, analyzed from space shows vent migration due to erosion. *Scientific Reports*, 15 (1). <https://doi.org/10.1038/s41598-025-95197-2>
- Archive Download. NASA. Available at: <https://firms.modaps.eosdis.nasa.gov/download/>
- Bjånes, A., De La Fuente, R., Mena, P. (2021). A deep learning ensemble model for wildfire susceptibility mapping. *Ecological Informatics*, 65, 101397. <https://doi.org/10.1016/j.ecoinf.2021.101397>
- Jones, M. W., Veraverbeke, S., Andela, N., Doerr, S. H., Kolden, C., Mataveli, G. et al. (2024). Global rise in forest fire emissions linked to climate change in the extratropics. *Science*, 386 (6719). <https://doi.org/10.1126/science.adl5889>
- Sharma, S. K., Aryal, J., Rajabifard, A. (2022). Remote Sensing and Meteorological Data Fusion in Predicting Bushfire Severity: A Case Study from Victoria, Australia. *Remote Sensing*, 14 (7), 1645. <https://doi.org/10.3390/rs14071645>
- Cao, G., Wei, X., Ye, J. (2024). Fireground Recognition and Spatio-Temporal Scalability Research Based on ICESat-2/ATLAS Vertical Structure Parameters. *Forests*, 15 (9), 1597. <https://doi.org/10.3390/f15091597>
- Liu, J., Wang, Y., Lu, Y., Zhao, P., Wang, S., Sun, Y., Luo, Y. (2024). Application of Remote Sensing and Explainable Artificial Intelligence (XAI) for Wildfire Occurrence Mapping in the Mountainous Region of Southwest China. *Remote Sensing*, 16 (19), 3602. <https://doi.org/10.3390/rs16193602>
- Dong, H., Wu, H., Sun, P., Ding, Y. (2022). Wildfire Prediction Model Based on Spatial and Temporal Characteristics: A Case Study of a Wildfire in Portugal's Montesinho Natural Park. *Sustainability*, 14 (16), 10107. <https://doi.org/10.3390/su141610107>
- Zacharakis, I., Tsihrintzis, V. A. (2023). Environmental Forest Fire Danger Rating Systems and Indices around the Globe: A Review. *Land*, 12 (1), 194. <https://doi.org/10.3390/land12010194>
- Di Giuseppe, F., Vitolo, C., Krzeminski, B., Barnard, C., Maciel, P., San-Miguel, J. (2020). Fire Weather Index: the skill provided by the European Centre for Medium-Range Weather Forecasts ensemble prediction system. *Natural Hazards and Earth System Sciences*, 20 (8), 2365–2378. <https://doi.org/10.5194/nhess-20-2365-2020>
- Eden, J. M., Krikken, F., Drobyshv, I. (2019). An empirical prediction approach for seasonal fire risk in the boreal forests. *International Journal of Climatology*, 40 (5), 2732–2744. <https://doi.org/10.1002/joc.6363>
- Arif, M., Alghamdi, K. K., Sahel, S. A., Alosaimi, S. O., Alsahafi, M. E., Alharthi, M. A., Arif, M. (2021). Role of Machine Learning Algorithms in Forest Fire Management: A Literature Review. *Journal of Robotics and Automation*, 5 (1). <https://doi.org/10.36959/673/372>
- Shmuel, A., Heifetz, E. (2022). Global Wildfire Susceptibility Mapping Based on Machine Learning Models. *Forests*, 13 (7), 1050. <https://doi.org/10.3390/f13071050>
- Ma, W., Feng, Z., Cheng, Z., Chen, S., Wang, F. (2020). Identifying Forest Fire Driving Factors and Related Impacts in China Using Random Forest Algorithm. *Forests*, 11 (5), 507. <https://doi.org/10.3390/f11050507>
- Zhang, L., Shi, C., Zhang, F. (2024). Predicting Forest Fire Area Growth Rate Using an Ensemble Algorithm. *Forests*, 15 (9), 1493. <https://doi.org/10.3390/f15091493>

20. Shmuel, A., Heifetz, E. (2023). Developing novel machine-learning-based fire weather indices. *Machine Learning: Science and Technology*, 4 (1), 15029. <https://doi.org/10.1088/2632-2153/acc008>
21. Alkan Akinci, H., Akinci, H., Zeybek, M. (2024). Comparison of diverse machine learning algorithms for forest fire susceptibility mapping in Antalya, Türkiye. *Advances in Space Research*, 74 (2), 647–667. <https://doi.org/10.1016/j.asr.2024.04.018>
22. Radke, D., Hessler, A., Ellsworth, D. (2019). FireCast: Leveraging Deep Learning to Predict Wildfire Spread. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 4575–4581. <https://doi.org/10.24963/ijcai.2019/636>
23. Wijikumar, P., Partheepan, S., Hassan, J., Sanati, F., Ray, B. R. (2024). Bushfire Severity Prediction Optimization: Feature Combination Study with XGBoost Regression and Satellite-Derived Data. *2024 17th International Conference on Security of Information and Networks (SIN)*, 1–7. <https://doi.org/10.1109/sin63213.2024.10871827>
24. Hong, H., Tsangaratos, P., Ilia, I., Liu, J., Zhu, A.-X., Xu, C. (2018). Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. The case of Dayu County, China. *Science of the Total Environment*, 630, 1044–1056. <https://doi.org/10.1016/j.scitotenv.2018.02.278>
25. Militino, A. F., Goyena, H., Pérez-Goya, U., Ugarte, M. D. (2024). Logistic regression versus XGBoost for detecting burned areas using satellite images. *Environmental and Ecological Statistics*, 31 (1), 57–77. <https://doi.org/10.1007/s10651-023-00590-7>
26. ukraine_geojson / UA_53_Poltavska.geojson. Available at: https://github.com/EugeneBorshch/ukraine_geojson/blob/master/UA_53_Poltavska.geojson
27. Esri. Available at: <https://www.esri.com/en-us/home>
28. The Weather API Built for Airlines, Marinas and Global Enterprises. Available at: <https://www.worldweatheronline.com>
29. Google Earth Engine. Available at: <https://earthengine.google.com>
30. Lundberg, S. M., Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. <https://doi.org/10.48550/arXiv.1705.07874>
31. Saito, T., Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *Plos One*, 10 (3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
32. Lan, Y.-D. (2017). A Hybrid Feature Selection Based on Mutual Information and Genetic Algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, 7 (1), 214. <https://doi.org/10.11591/ijeecs.v7.i1.pp214-225>
33. Saha, M. V., Scanlon, T. M., D'Odorico, P. (2018). Climate seasonality as an essential predictor of global fire activity. *Global Ecology and Biogeography*, 28 (2), 198–210. <https://doi.org/10.1111/geb.12836>
34. Varela, V., Vlachogiannis, D., Sfetsos, A., Karozis, S., Politi, N., Giroud, F. (2019). Projection of Forest Fire Danger due to Climate Change in the French Mediterranean Region. *Sustainability*, 11 (16), 4284. <https://doi.org/10.3390/su11164284>
35. Andronov, V., Pospelov, B., Rybka, Evgeniy, Skliarov, S. (2017). Examining the learning fire detectors under real conditions of application. *Eastern-European Journal of Enterprise Technologies*, 3 (9 (87)), 53–59. <https://doi.org/10.15587/1729-4061.2017.101985>
36. Pospelov, B., Andronov, V., Rybka, Evgeniy, Skliarov, S. (2017). Design of fire detectors capable of self-adjusting by ignition. *Eastern-European Journal of Enterprise Technologies*, 4 (9 (88)), 53–59. <https://doi.org/10.15587/1729-4061.2017.108448>