Wildfire Prediction and Detection using Random Forest and Different Color Models

Ananya Shahdeo[1], Aditi Shahdeo[1], Prakruthi S Reddy[1], Chaitra K[2]

[1]Student, Dept. of Computer Science and Engineering, Bangalore Institute of Technology, Bengaluru, Karnataka, India
[2]Assistant Professor, Dept. of Computer Science and Engineering, Bangalore Institute of Technology, Bengaluru, Karnataka, India

---***---

Abstract - *In the recent years, wildfires have immerged to be one of the major threats to the world's environment and has caused tremendous amount of destruction. This paper proposes a solution which can predict the possibility of the wildfire based on factors such as meteorological, topographical, vegetation factors and various fire weather indices that influence the occurrence of wildfire to a large extent. Prediction is carried out using a machine learning approach. This prediction result determines the possibility of the wildfire. This result can be used to monitor the areas that are predicted to have a risk of wildfire occurrence which helps in implementing methods that can be used to minimize the effects of wildfire. As an additional feature the proposed solution also includes the detection of these wildfires using image processing approach wherein satellite image map spaces of the respective area are extracted for detection. This detection feature is used to test the accuracy of the prediction based on previous fire data and the corresponding map space images.*

***Key Words*: Meteorological, Topographical, Fire weather indices, Prediction, Machine learning, Image Processing, Satellite image map space.**

1. INTRODUCTION

Wildfires are fires that burn out of control in a natural area, like a forest area and grasslands. Wildfires pose a serious threat to the world's environment as forests account for more than 31% of the world's land surface and contribute to the continuity of ecological balance and play a paramount role in environmental sustainability. They spread very quickly, and the destructive nature of a wildfire in a forest is phenomenal. If a wildfire strikes, it kills and displaces wildlife, leads to loss of vegetation, alters water cycles and soil fertility, endangers the lives and livelihoods of local communities, and also releases harmful pollutants that include particulate matter and toxic gases that may contribute to global warming. These wildfires remain unidentified until a large area has been affected and in most cases the source is unknown. Wildfires can be characterized in terms of the cause of ignition, their physical properties, the combustible material present, and the effect of weather on the fire. The risk of wildfires increases in extremely dry conditions, such as drought, and during high winds. Therefore, prediction and detection of these wildfires can help fire prevention, fire suppression and forest management. This may help in mitigating the disastrous effects of the wildfire.

The objective of the proposed solution is to build a system using advanced algorithms that provide great ease in wildfire prediction and detection. This can be implemented by forest fire departments to accomplish a safer environment for wildlife and promote biodiversity.

1.1 EXISTING SYSTEM

In prediction, the geographic information collected by post-fire field survey aims to examine the causes and effects of forest fires and are usually recorded in the form of a postal address.

Limitations:

- The data lacks precise ignition point information.
- Countries with poor infrastructure are often unable to conduct their own field surveys which make it difficult to obtain basic historical data on previous fire sources, thus restricting their accurate prediction.

In detection, sensor technology has been widely used in detection, usually depending on sensing physical parameters and chemical parameters.

Limitations:

- It is difficult to apply these systems in large open areas for a variety of reasons such as high cost, energy usage by the sensors.

- High false alarms rate
- Large response time

1.2 PROBLEM STATEMENT

To develop a system that predicts wildfires using random forest approach and detect this fire in the corresponding map space image using HSV color model.
Therefore, the proposed solution is designed to:

- Train a prediction model using Random forest algorithm.
- Predict the possibility of wildfire of the given set of attributes.
- Detect fire in the same map space confirming the presence of wildfire in the predicted area.

2. MATERIALS AND METHODS

2.1 Data Collection and Data description

For prediction, data is collected from the MODIS Active Fire data which is provided by NASA's FIRMS (Fire Information and Resource Management System). Moderate Resolution Imaging Spectroradiometer (MODIS) is an instrument that operates on both the Terra and Aqua spacecraft. The data products from MODIS observations describe features of the land, oceans and the atmosphere that can be used for studies of processes and trends on local to global scales. This data includes various different sets of data from which the final data is manually extracted with the list of attributes that largely influence the occurrence of a wildfire. For detection, satellite map space images are collected for the same data corresponding to the prediction dataset. Table-1 lists out the attributes used for prediction of wildfire.

Meteorological factors include the factors that are concerned with the processes and phenomenon of the atmosphere like relative humidity, rain, temperature. Topographical factors include the factors related to the arrangement or accurate representation of physical distribution of features of an area like land surface temperature (LST), burnt area. Vegetation factors like Normalized Difference Vegetation Index (NDVI) that can be used to analyse remote sensing measurements that can be used for assessing whether the target being observed contains live green vegetation. The components of Fire Weather Index (FWI) are meteorologically based components used worldwide to estimate fire danger. It consists of different components that account for the effects of fuel moisture and wind on fire behavior and spread. Calculation of these components is based on consecutive daily observations of temperature, relative humidity, wind speed, and 24-hour precipitation. They include Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), and Initial Spread Index (ISI).

Table -1: List of attributes used for Prediction

Attributes	Explanation
FFMC	The Fine Fuel Moisture Code is an indicator of the relative ease of ignition and the flammability of fine fuel.
DMC	The Duff Moisture Code is an indication of fuel consumption in moderate duff layers and medium-size woody material.
DC	The Drought Code is an indicator of seasonal drought effects on forest fuels and the amount of smoldering in deep duff layers and large logs.
ISI	The Initial Spread Index combines the effects of wind and the FFMC on rate of spread without the influence of variable quantities of fuel.
Temperature	It is the temperature in ° C
Rh	Relative humidity (RH) is the ratio of the amount of moisture in the air to the amount of moisture necessary to saturate the air at the same temperature and pressure
Wind	Wind increases the supply of oxygen, which results in the fire burning more rapidly. It also removes the surface fuel moisture, which increases the drying of the fuel
Rain	Rainfall in open measured in mm
Area	Area covered in hectares
NDVI	The normalized difference vegetation index is a simple graphical indicator that can be used to analyze remote sensing measurements, assessing whether or not the target being observed contains live green vegetation.
LST	Land Surface Temperature (LST) is the radiative skin temperature of the land derived from solar radiation.
Burnt Area	This index highlights burned land in the red to near-infrared spectrum, by emphasizing the charcoal signal in post-fire images.

2.2 Machine Learning Algorithms

Machine learning algorithms are programs that can learn from data and improve from experience, without human intervention. Learning tasks may include learning the function that maps the input to the output or learning the hidden structure in unlabeled data. Supervised learning uses labelled training data to learn the mapping function that turns input variables (X) into the output variable (Y). There are different machine learning techniques like Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbor, Decision Tree, Random Forest, Gaussian Naïve

Bayes and Support Vector Machine used for prediction. Here, the accuracy of these algorithms is measured based on the data using ten-fold method, as a result of which Random forest gives the best performance with highest accuracy range. Therefore, Random Forest algorithm is used to build the proposed solution.

2.2.1 Random Forest Algorithm

Random Forest algorithm is used to train and classify the prediction model. Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and output the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Algorithm:
Input: Forest size (Number of decision trees)
Output: Predicted class (1 for fire predicted and 0 for no fire predicted)
Step 1 – Selection of random sample from a given dataset (Bootstrapping) of a given size.
Step 2 – The algorithm constructs a set decision trees (random forest) for bootstrap dataset based on the size of forest. Here the model is trained.
Step 3 – For every test instance, voting will be performed on each of the decision tree to generate the prediction result.
Step 4 – At last, select the most voted prediction result as the final prediction result.

2.3 Image Processing

Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. It is a type of signal processing in which input is an image and output may be image or features associated with that image. In the paper, this technique is used to detect the region of wildfire that are represented as hotspots in the satellite image map space of the area where the possibility of wildfire is predicted.

HSV color space stands for Hue, Saturation, and Value (or brightness). We have an RGB (Red-Green-Blue) image with red dots called hotspots that determine the wildfire as input. The RGB image is converted to HSV color space and is used to detect the hotspots in the map space. The upper and lower limit is set using the range-detector script in the imutils library for masking and extracting the hotspots in the hsv image and then these values are put into a numpy array. The mask represents a specific part of the image that contains the hotspots. Finally, the areas of red to give the detection result are counted.

3. IMPLEMENTATION

A. COMPARISION

Different machine algorithms are considered in order to find the best prediction algorithm that can be used in the proposed solution. These algorithms include:

Logistic Regression is an approach to learning functions of the form f : X →Y, or P(Y|X) in the case where Y is discrete-valued, and X = <X_1 ...X_n>is any vector containing discrete or continuous variables. It assumes a parametric form for the distribution P(Y|X), then directly estimates its parameters from the training data. The parametric model assumed by Logistic Regression in the case where Y is boolean is:

$$P(Y=1|X)=\frac{1}{1+\exp(w_0+\sum_{i=1}^{n} w_i X_i)}$$

and

$$P(Y=0|X)=\frac{\exp(w_0+\sum_{i=1}^{n} w_i X_i)}{1+\exp(w_0+\sum_{i=1}^{n} w_i X_i)}$$

Linear Discriminant Analysis (LDA) is a supervised classification technique that is considered a part of crafting competitive machine learning models. LDA focuses primarily on projecting the features in higher dimension space to lower dimensions. You can achieve this in three steps:

- The separability between classes which is the distance between the mean of different classes. This is called the between-class variance is calculated.
- Then, the distance between the mean and sample of each class, also called the within-class variance is calculated.
- Finally, the lower-dimensional space which maximizes the between-class variance and minimizes the within-class variance is calculated.

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new data and available data and put the new data instance into the category that is most similar to the available categories. The K-NN working can be explained based on the below algorithm:

- Select the number K of the neighbors.
- Calculate the Euclidean distance of K number of neighbors.
- Take the K nearest neighbors as per the calculated Euclidean distance.
- Among these k neighbors, count the number of the data points in each category.
- Assign the new data points to that category for which the number of the neighbor is maximum.

Decision Tree is a supervised learning technique that can be used for both classification and regression problems, but mostly it is preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Gaussian Naïve Bayes is an algorithm having a Probabilistic Approach. It involves prior and posterior probability calculation of the classes in the dataset and the test data given a class respectively.

$$Prior\ Probability(c) = \frac{No.of\ instances\ of\ class\ c}{Total\ No.of\ instances\ in\ the\ dataset}$$

$$Posterior\ Probability(x \mid c) = P(x_1 \mid c) * P(x_2 \mid c) * P(x_3 \mid c) * \ldots * P(x_n \mid c)$$

This is given by the probability obtained from Gaussian (Normal) Distribution.

$$P(x_i \mid c) = \frac{1}{\sqrt{2 * \pi * sigma_{x_i,c}^2}} * \exp\left(-\frac{\left(x_i - mean_{x_i,c}\right)^2}{2 * sigma_{x_i,c}^2}\right)$$

Finally, the conditional probability of each class given an instance (test instance) is calculated using Bayes Theorem

$$P(c_i \mid x) = \frac{P(x \mid c_i) * P(c_i)}{\sum_j P(x \mid c_j) * P(c_j)}$$

Support Vector Machine or SVM is one of the supervised learning algorithms, which is used for classification as well as regression problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

In this paper, ten-fold method is used on the train data wherein the dataset is divided into 10 sub-datasets and the above mentioned algorithms are tested on each of the subsets and the following accuracy was observed:

Logistic Regression: 75.45%
Linear Discriminant Analysis: 75.01%
K-Nearest Neighbor: 73.27%
Decision Tree: 68.23%
Random Forest: 77.12%
Gaussian Naïve Bayes: 75.36%
Support Vector Machine: 76.06%

It can be inferred by the above accuracies that random forest gives the best performance. Therefore, in this paper the proposed solution is designed using Random Forest Algorithm.

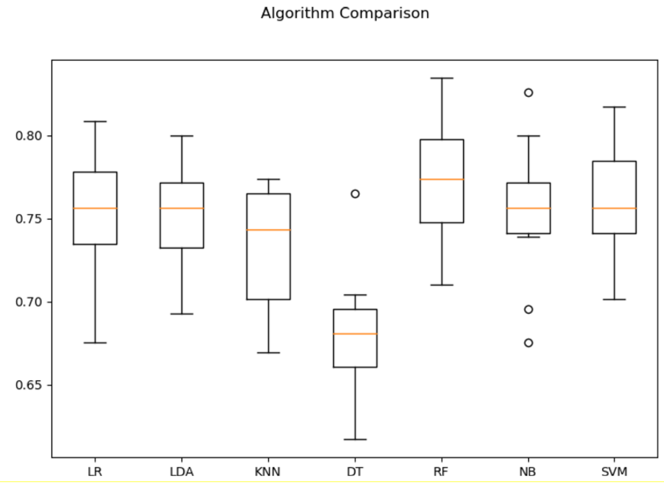


Fig 1: Result of Algorithm Comparison

B. WORKING OF THE SYSTEM

For prediction, Random forest algorithm is used. A random forest is a model made of many decision trees. The model has two of the key components which are (i) random sampling of training data points when building trees and (ii) random subsets of features considered when splitting nodes. When training, each tree in a random forest learns from a random sample of the data points. The samples are drawn with replacement, known as bootstrapping, which means that some samples will be used multiple times in a single tree. The idea is that by training each tree on different samples, although each tree might have high variance with respect to a particular set of the training data, overall, the entire forest will have lower variance but not at the cost of increasing the bias. During testing, predictions are calculated as the average of predictions of each decision tree. This procedure of training each individual learner on different bootstrapped subsets of the data and then averaging the predictions is known as bootstrap aggregating.

After bootstrapping, for each bootstrap sample, a decision tree is constructed, and all the decision trees combined forms the forest. To build the decision tree entropy (measure of uncertainty) and information gain for each attribute is calculated. The attribute that maximizes the information gain, which in turn minimizes the entropy and best splits the dataset into groups for effective classification is selected and is determined as the root node of the tree. This process is repeated for remaining set of attributes until the entire decision tree is constructed. At the end of the training process, it is observed that 'n' number of decision trees for a given forest size 'n'.

In the classification phase of random forest algorithm, the votes from different decision trees to decide the final class of the test object are aggregated. Each test instance is classified to one output class for each decision tree in the forest. Each classification output is known as a vote. The final classification is done based on majority voting among all the votes of each tree. For each test instance the classification is done based on majority voting to determine the output class (0 for not predicted or 1 for predicted) for the given test instance.

For detection, the satellite image map space of the area corresponding to the area considered for prediction is extracted. The extracted image is in the RGB color space which is then converted to HSV color space. HSV is an alternative representation of the RGB color model that is designed to more closely align with the way human vision perceives to the color -making attributes. The aim is to identify the hotspots in the map space. These hotspots represent the active fire region in the map space. The lower and upper limit is set to ([161, 155, 84]) and ([179, 255, 255]) respectively for masking the hotspots as red and the region other than the hotspots as black . Then the number of hotspots is counted using the imultis grab contour. If the number of hotspots exceeds the set threshold value then, the result of detection is given as fire detected otherwise as not detected.

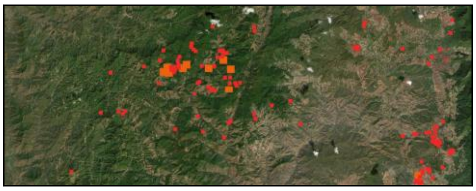


Fig 2: Satellite map space image

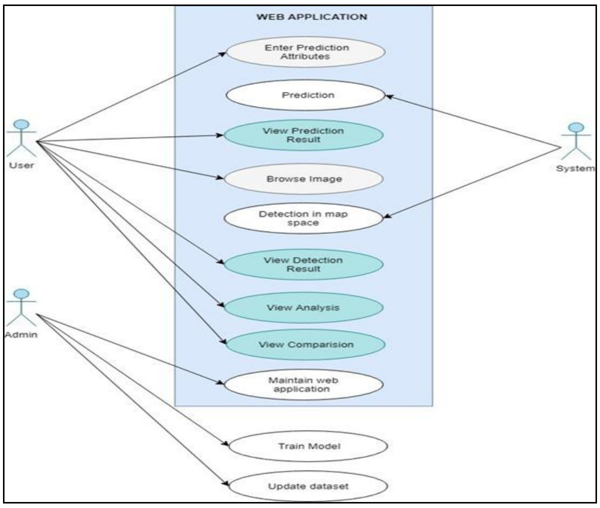


Fig 3: Use-Case diagram of the system

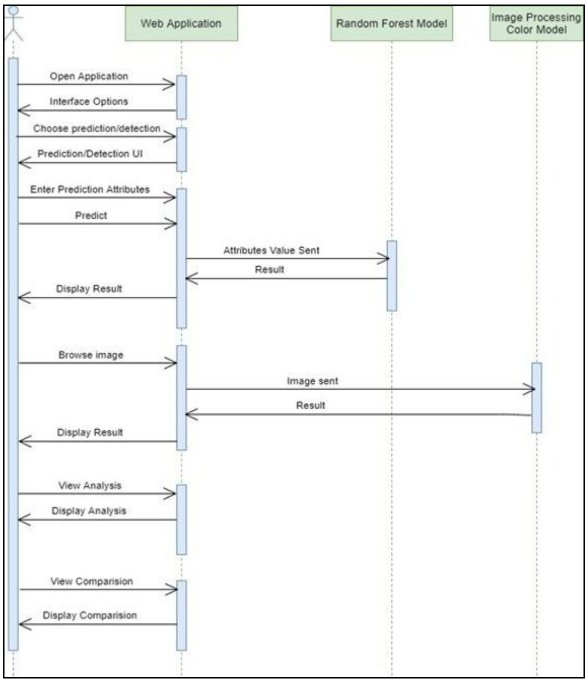
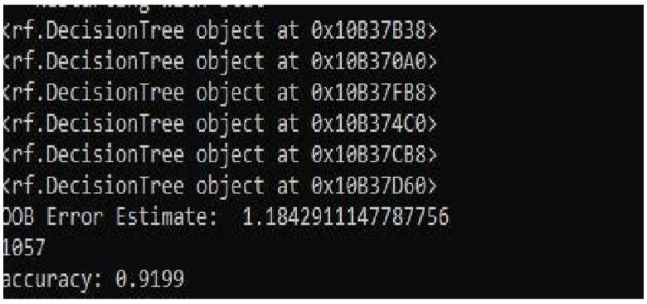


Fig 4: Sequence diagram of the system

C. RESULTS

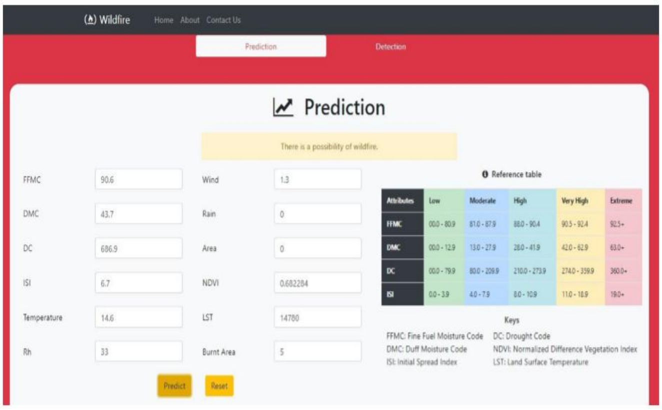
The entire dataset is tested against the model and the actual values are compared with the predicted results. The model gives an accuracy of 91%.



```
rf.DecisionTree object at 0x10B37B38>
rf.DecisionTree object at 0x10B370A0>
rf.DecisionTree object at 0x10B37FB8>
rf.DecisionTree object at 0x10B374C0>
rf.DecisionTree object at 0x10B37CB8>
rf.DecisionTree object at 0x10B37D60>
OOB Error Estimate:  1.1842911147787756
1057
accuracy: 0.9199
```

Fig 5: Accuracy of the algorithm

A UI is created using Angular JavaScript and FLASK API is used to run the prediction and detection system for real time data. This web application can be used by the user where the attribute values are entered, and the output can be predicted. Also, the map space image can be uploaded to detect if the wildfire has occurred or not.



Prediction

There is a possibility of wildfire.

FFMC	90.6	Wind	1.3
DMC	43.7	Rain	0
DC	686.9	Area	0
ISI	6.7	NDVI	0.682284
Temperature	14.6	LST	14780
Rh	33	Burnt Area	5

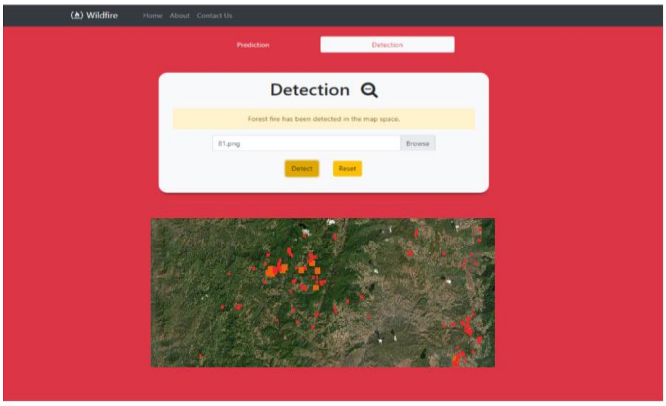
Reference table

Attributes	Low	Moderate	High	Very High	Extreme
FFMC	00.0 - 80.9	81.0 - 87.9	88.0 - 90.4	90.5 - 92.4	92.5+
DMC	00.0 - 12.9	13.0 - 27.9	28.0 - 41.9	42.0 - 62.9	63.0+
DC	00.0 - 79.9	80.0 - 209.9	210.0 - 273.9	274.0 - 359.9	360.0+
ISI	0.0 - 3.9	4.0 - 7.9	8.0 - 10.9	11.0 - 18.9	19.0+

Keys

FFMC: Fine Fuel Moisture Code
DMC: Duff Moisture Code
ISI: Initial Spread Index
DC: Drought Code
NDVI: Normalized Difference Vegetation Index
LST: Land Surface Temperature

Fig 6: Prediction Result displayed in UI



Detection

Forest fire has been detected in the map space.

81.png

Fig 7: Detection Result displayed in UI

4. CONCLUSIONS

Mapping the prediction of wildfire susceptibility is an essential component of emergency land management, wildfire prevention, the mitigation of fire impacts by on-time responses and recovery management. Wildfire susceptibility maps have often been used to prioritize investments in the prevention of this hazard. In this paper, Random forest is used because it gives the highest accuracy among different prediction algorithms and it has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing. It also runs efficiently on large datasets and can handle thousands of input variables without variable deletion. It gives estimates of what variables are important in the classification.

To get the corresponding map space image of the instances in the prediction data set, the coordinates need to be mapped manually. As future work ,this can be improved by automating this process.

ACKNOWLEDGEMENT

We are most humbled to mention the enthusiastic influence provided by our guide Prof. Chaitra K., Assistant Professor for her ideas, time to time suggestions and co- operation showed during the venture and helping make our project a success.

REFERENCES

[1] Babu, Suresh & Kabdulova, G & Kabzhanova, G. “Developing the Forest Fire Danger Index for the Country Kazakhstan by Using Geospatial Techniques”,2019.

[2] Chul-Hee Lim, You Seung Kim, Myungsoo Won, Sea Jin Kim & Woo-Kyun Lee Geomatics, Natural Hazards and Risk “Can satellite-based data substitute for surveyed data to predict the spatial probability of forest fire? A geostatistical approach to forest fire in the Republic of Korea”, 10:1, 719-739 ,2019.

[3] Dieu Tien Buia, Nhat-Duc Hoangc, Pijush Samuid, “Spatial pattern analysis and prediction of forest fire using new machine learning approach of Multivariate Adaptive Regression Splines and Differential Flower Pollination optimization: A case study at Lao Cai province (Viet Nam)”,Journal of Environmental Management, Volume 237, Pages 476-487 , 1 May 2019.

[4] Kim, S.; Lim, C.-H.; Kim, G.; Lee, J.; Geiger, T.; Rahmati, O.; Son, Y.; Lee, W.-K, "Multi-temporal analysis of forest fire probability using socio-economic and environmental variables", Remote Sens. 2019.

[5] Haoyuan Hong, Paraskevas Tsangaratos,Ioanna Ilia,Junzhi Liu, A-Xing Zhu , Chong Xu, "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 ,Volume 630,Pages 1044-1056,15 July 2018.