SINERGI Vol. 23, No. 3, October 2019: 184-190
Accredited **SINTA 2** by KEMENRISTEKDIKTI, Decree No: 10/E/KPT/2019
http://publikasi.mercubuana.ac.id/index.php/sinergi
http://doi.org/10.22441/sinergi.2019.3.002 .

THE IDENTIFICATION OF DETERMINANT PARAMETER IN FOREST FIRE BASED ON FEATURE SELECTION ALGORITHMS

Devi Fitrianah[1] Hisyam Fahmi[2]
[1]Department of Informatics, Faculty of Computer Science, Universitas Mercu Buana
Jl. Raya Meruya Selatan, Kembangan, Jakarta 11650, Indonesia
[2]Department of Maths, Universitas Islam Negeri Maulana Ibrahim Malang
Jl. Gajayana No.50, Dinoyo, Lowokwaru, Malang 65144, Indonesia
Email: devi.fitrianah@mercubuana.ac.id hisyam.fahmi@uin-malang.ac.id

Abstract -- *This research conducts studies of the use of the Sequential Forward Floating Selection (SFFS) Algorithm and Sequential Backward Floating Selection (SBFS) Algorithm as the feature selection algorithms in the Forest Fire case study. With the supporting data that become the features of the forest fire case, we obtained information regarding the kinds of features that are very significant and influential in the event of a forest fire. Data used are weather data and land coverage of each area where the forest fire occurs. Based on the existing data, ten features were included in selecting the features using both feature selection methods. The result of the Sequential Forward Floating Selection method shows that earth surface temperature is the most significant and influential feature in regards to forest fire, while, based on the result of the Sequential Backward Feature Selection method, cloud coverage, is the most significant. Referring to the results from a total of 100 tests, the average accuracy of the Sequential Forward Floating Selection method is 96.23%. It surpassed the 82.41% average accuracy percentage of the Sequential Backward Floating Selection method.*

Keywords: *Forest fire; Data mining; Feature selection; SFFS Algorithm; SBFS Algorithm*

Received: May 23, 2019 *Revised: September 25, 2019* *Accepted: October 6, 2019*

INTRODUCTION

Forest fire is a global issue that has become the attention of several nations, and one of them is Indonesia. The forest fire disaster that occurred in July 2015 is the worst in a decade of forest fire history in Indonesia. According to LAPAN (Indonesian National Institute of Aeronautics and Space) Head of Environment and Disaster Mitigation, there are around 2.1 million hectares of burnt land and forest in 2015 (Lelloisima, 2015). According to existing forest fire data, Indonesia, with about 680.000 hectares of burnt land and forest per year in 2009, has officially become one of the nations with the highest deforestation rate in the world (FAO, 2010).

The Indonesian government has made some efforts to restore the burnt forest area, including the rehabilitation of forest and land (Kusmana et al., 2004). However, based on the condition that happened, it is not enough to rely on the rehabilitation effort to restore the burnt land and forest caused by the fire.

Many efforts could be made to minimize and prevent forest fire, including the studies done in some nations. One of the studies done in order to avoid the occurrence of forest fire is by creating a distribution pattern model of a forest fire that happens spatially and temporally, or in other words, mapping the forest fire (Orozco et al., 2012). Another study is done to predict, based on spatial and temporal algorithms, to determine forest fire that will occur (Zhao et al., 2011). Efforts to detect forest fire early has been made by using a machine learning approach, as an example, the one done by Zhao et al. (2011). Almost all real-time detection approaches are using remote sensing technology in the early determination forest fire area (Alkhatib, 2014; Zhao et al., 2015).

Furthermore, some studies are focusing on the analysis of the most influential features in the occurrence of a forest fire. There are three very influential factors (features) in a forest fire, vegetation land coverage, weather and climate, and topography (Wagtendonk & Lutz, 2007). By knowing forest fire cause parameters, there is hope that there will be a contribution to the advanced study in mapping forest fire-prone areas based on land vegetation coverage, weather, and topography. From some of these parameters, the selection of the most influential feature can be made

according to its spatiality and temporality. Thus, this research will perform an experimental study to some of the previously mentioned parameters in several regions and on a certain time period. It is not easy to determine which parameters are influential towards forest fire, moreover, if they are analyzed based on their spatiality and temporality. Therefore, research regarding the data-driven method will be done. SFSS algorithm (*Sequential Forward Floating Selection*) and SBSF algorithm (*Sequential Backward Floating Selection*) are the approaches used because these algorithms are applicable in selecting the parameter or influential feature. These algorithms once have been used for selecting marine oceanography parameter features, which has become the reference of fish catch potential area with the significant results (Fitrianah, 2015). SFSS and SBFS algorithms are algorithms commonly used to find the optimal features set using inclusion, which means combining new features with a group of existing features then continued with performing exclusion. These processes then will exclude the worst features from the set one by one gradually. SFFS algorithm is an advancement of the SFS algorithm, with exclusion being the only difference between the two algorithms. In the SFFS algorithm, the exclusion of a feature that already is a part of a feature set and can be re-include once there is a new feature will be used as the comparison (Pudil, Novovičová, & Kittler, 1994).

MATERIAL AND METHOD

In this section, we will explain the related study in our research. We will discuss forest fire, feature selection, and classification. The method used in this paper will be described in this section, as well.

Forest Fire Study

Many researchers all over the world have performed researches regarding forest fire. Especially in relation to nations that have forest extents. Almost most studies concerning forest fire utilize sensors from remote sensing technology. Umamaheshwaran proposed an image mining method by using images from Meteosat-SEVIRI sensors that produce a linear model of a forest fire with existing vegetation and wind direction (Umamaheshwaran, Bijker, & Stein, 2007). Other studies studied variables that affected forest fire or influential parameters in the process of a forest fire, thus enabling the formulation of their relationship with the types of a forest fire. These influential parameters are the time of the fire, land coverage, height of the forest area, ground slant, and forest fire statistic. This model has been validated by using the data obtained from NOAA-AVHRR and Terra MODIS satellites (Hernández-Leal et al., 2008). Aside from these parameters, another thing that should be considered in determining forest fire prediction is the size of the grid of the image used in existing weather observation (Khabarov et al., 2008).

Another research concerning the variables related to forest fire is finding out how to determine the relationship between land coverage/vegetation with occurring forest fire (Tanase & Gitas, 2008). This study successfully described the potential vegetation types of forest fire by using the help of images from remote sensing satellites. Similar research identifies the relationship between forest fire and its vegetation utilizing the help of images from the MODIS satellite (Biswas, Lasko, & Vadrevu, 2015). Further study on forest fire already reach the impact of forest fire utilizing the satellite data over the Northeast region in India, and it concludes a significant correlation between forest fire occurrences and variations in the aerosol concentrations over the study region (Badarinath, Kharol, & Chand, 2007).

Feature Selection

Feature selection is an activity done in the pre-process aims to achieve and select influential features in modeling and data analysis. Generally, there are two major groups in feature selection: ranking selection and subset selection (Wang et al., 2015).

Ranking selection provides order to features which then exclude substandard features. There are some ways done in the ranking selection method. Some of them are regression, correlation, mutual information, etc.

Subset selection is a selection method used to find a set of features that are considered as the optimal features. There are three types of methods commonly used in subset selection: wrapper type selection, filter type selection, and embedded type selection (Kohavi & John, 1997). Wrapper type selection makes the selection simultaneously with modeling implementation. The method proposed by Kohavi and John (Kohavi & John, 1997) proved to be great and able to overcome the issue of feature selection well (Guyon & Elisseeff, 2003). This selection type uses criteria that utilize classification average from the classification method used. Examples of its algorithms are Sequential Forward Selection (from one into many features), Sequential Backward Selection (from many features into

one), Sequential Floating (can be either backward or forward), Generic Algorithm, Greedy Search, Hill Climbing, Simulated Annealing, and many others.

As stated in the previous section, to get the optimal feature, we can use wrapper type selection. To obtain a series of optimal features from several existing feature inputs, Sequential Forward Floating Selection, or Sequential Backward Floating Selection methods are two usable approaches (Pudil et al., 1994). The SFFS algorithm consists of three main steps; the first step is to select the feature from the set of available measurements to form a set of the most significant feature with respect to the set X_k is added to X_k, therefore $X_{k+1} = X_k+x_{k+1}$. Second, we do the conditional exclusion by finding the least significant feature in the set X_{k+1}. The third is continuing the conditional exclusion. The SFFS is continued with step 1 until a feature set of cardinalities two is obtained. The SBFS is the opposite of the SFFS. There are also three main steps in SBFS; first, exclude the feature from the current set to form a reduced feature set. The least significant feature is removed from the set. Second, among the excluded features, we find the most significant feature and continue until no feature left. The last step is to continue the conditional inclusion until the new enlarged set is formed.

These feature selection methods have been proven to produce high accuracy results in determining the most relevant feature based on certain criteria (Fitrianah, 2015).

Classification Model

Classification Model is a technique in data mining to group instance data into the appropriate classes. In its implementation, there are two processes, namely, sample data training process and testing process. The sample data training process is the one named as a supervised learning technique. As training, classification is used to build a model from given data.

Classification can be used for a variety of issue approaches, such as IR (Information Retrieval) (Joshi & Nigam, 2011), Geography and Remote Sensing (Zhong et al., 2014), Web Technology (Ali, Shamsuddin, & Ismail, 2012), and others. In the process, training and testing require variable, which is commonly called feature. This feature enables exploration done to give optimum results. There are several well-known basic classification algorithms including algorithms with probabilistic basis (Ali, Shamsuddin, & Ismail, 2012) such as Naive Bayes Classifier and Hidden Markov Model; information gained based classification (Decision Tree) (Xiaowei, 2014; Reddy et al., 1994), kernel-based (SVM) (Chuan-xu, Yun, & Zuo-yong, 2008; Changhui et al., 2010) and others.

Method

The research methodology is divided into several phases, as can be seen in Fig 1.

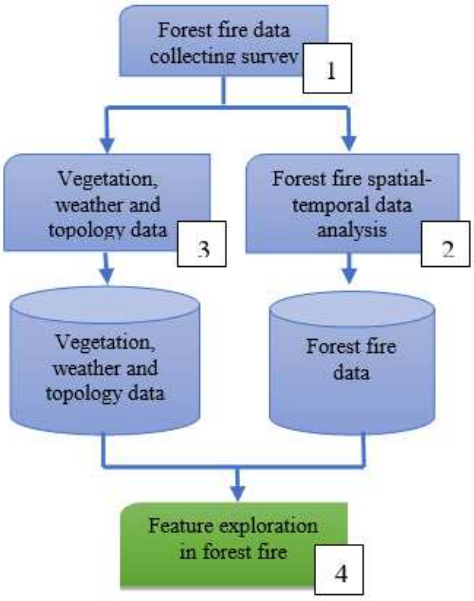

Figure 1. Block Diagram of research methodology

Forest Fire Data Collecting Survey

This activity collects forest fire data. Data required are spatial and temporal data. Forest fire data used extent ten years from 2005 to 2015. Riau province and Jambi province are selected as the research locations.

Spatial and Temporal Analysis of Forest Fire Data

Obtained forest fire data will be analyzed spatially and temporally and is adjusted with data structure used in a storage data. Furthermore, these data should contain information about the location and period of the occurred forest fire. Besides, the next analysis will require time period data. The result from this stage is the formation of a database about a forest fire that matches the desirable data structure.

Vegetation, Weather, and Topology Data Analysis

After obtaining forest fire data spatially and temporally, the next stage is to obtain relevant parameter data to forest fire, which are

vegetation data of the forest fire area, weather data, and topology data. After obtaining the relevant data, the next stage is to analyze those data based on the land coverage area and its time. The analysis will be done using the data integration approach, which is a hybrid data integration approach (Fitrianah & Wasito, 2012). It is hoped that this analysis will acquire information about parameter data of forest fire area. The results of this stage are the formation of a database regarding forest fire information and related parameter that matches its spatiality and temporality.

The implemented method in feature selection study is SFFS, which is a bottom-to-top search procedure by excluding features of SFS method's result in the beginning and followed by a series of more significant conditional inclusions afterward. The followings are the steps of the SFFS:

1. *Inclusion*: use the basic SFS method, select and input the most significant features in accordance with the predetermined feature set.
2. *Conditional exclusion*: find the least significant feature in the predetermined feature set, exclude the least significant feature from the new feature set.
3. *Continuation of conditional exclusion*: continue selecting the least significant feature until there is a no more least significant feature in the feature set.

The feature selection method will be conducted, as seen in Fig 2.

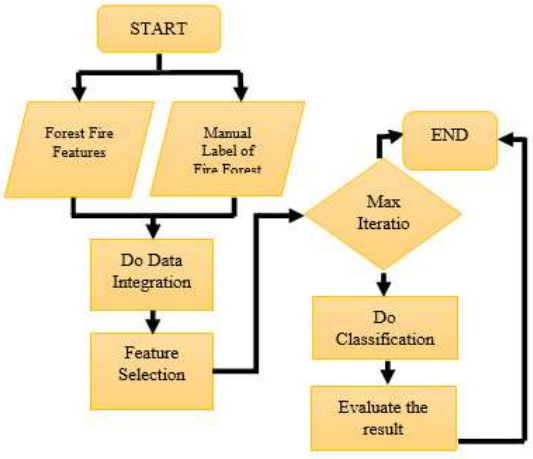

Figure 2. The framework of feature selection

Fig 2. explains the process of feature selection. The process done is determining the feature of forest fire following the forest fire manual label. After the data is collected, the next step is the integration of both data. After integrating the data, we do the feature selection using the SFFS algorithm. The result of feature selection using the SFFS method is a feature set that selectively more important than in the previous set. Then the maximum iteration is evaluated. When it fulfilled the maximum iteration, the result is then classified. K-Nearest Neighbor is used as the classification method.

RESULTS AND DISCUSSION

Data

There are three types of data used in this research. The first data is forest fire data, the second is weather data, and the third is land coverage data. Data used is forest fire data in the year of 2011. The limitation of the year 2011 is mostly because of the availability of other forest fire data. In this case, island coverage data. As for the vegetation data or land coverage, the authors obtained the data from the WebGIS Kehutanan (Forestry GIS Web), which is an official website of the Forestry Department of the Republic of Indonesia. There are 12 types of land coverage, according to the website. They are a secondary dryland forest, primary dryland forest, bush-mixed farmland, plantation, fishpond, paddy field, dry land field, bush/grove, bush/grove/swamp, industrial plants forest, and savanna. Weather data used are Temperature in degree Celsius, Weather Code, Precipitation mm, Humidity, Cloud Cover, Heat Index, Dew Point, Wind Speed, and Wind Gust.

Forest Fire Feature Exploration

Forest fire feature exploration is done using two algorithm approaches, SFFS and SBFS. The selection result by using the SFFS algorithm with ten features can be seen in Table 1.

There are features and scores in the chart. Feature states that those four features are the most influential features out of 19 presented features. These four features are feature number 4 - Precipitation or Air Condensation, feature number 10 - Wind Gust, feature number 9 - Wind Speed, and feature number 2 - Temperature.

Table 1. The score of the feature selection result using the SFFS

Feature	4 (PrecipMM)	10 (Wind GustKmph)	9 (Wind SpeedKmph)	2 (TempC)
Score	91.0417	93.75	96.25	97.2917

The scores underneath each feature state the classification score from each feature set, which means they show the level of influence from related features by the occurrence of a forest fire. The scores are generated from the Recall score of the features. Based on the classification measure, we utilized the Unweighted Average Recall (UAR) = mean (R1, R2), where R1 recalls from class 1, and R2 is the recall from class 2. The recall is based on the confusion matrix, as illustrated in Table 2.

Table 2. Confusion Matrix

	Class 1 Predicted	**Class 2 Predicted**
Class 1 Actual	True-Positive	False-Negative
Class 2 Actual	False-Positive	True-Negative

From the confusion matrix, we can calculate the recall score using the Equation (1) as follows:

$$Recal = \frac{TP}{TP + FN} \quad (1)$$

where:

TP = True Positive
FN = False Negative

From the score, we know the number of correctly classified positive examples divided by the total number of positive examples.

Based on Table 1, it can be seen that the most influential feature is earth surface temperature with a score of **97, 2917**. The second most influential feature is wind speed, with a score of 96, 25. The third is wind gust feature with a score of 93, 75, and the last is air condensation (precipitation) with a score of 91, 0417.

While the SFFS result shows that there are four very influential features in the occurrence of a forest fire, the SBFS feature selection method produced five features that affect forest fire occurrence. The result of the feature selection method using the SBFS method can be seen in Table 3.

Table 3. The Score of The Feature Selection Using the SBFS

Feature	**4 (PrecipMM)**	**5 (Humidity)**	**6 (CloudCover)**	**7 (HeatIndex)**	**8 (DewPoint)**
Score	81.0417	82.7083	88.75	86.6667	86.6667

From Table 3, based on the SBFS method, there are five very influential features that affect forest fire occurrence. The most influential one is the Cloud Cover feature or **cloud coverage**, with a score of **88, 75**.

To obtain accuracy, the overall selected features of each method are validated. The average feature selection from each method is made in 100 experiments, as shown in Table 4.

Table 4. Average accuracy for SFFS and SBFS

Method	**Average accuracy value (%)**
SFFS	96.63
SBFS	82.41

From Table 4, the average accuracy value from 100 experiments using the SFFS method is better than the SBFS method, which is 96, 63%. This also concludes that between the two methods, SFFS is better than the SBFS method regarding our dataset. This means that the feature selection process using the SFFS method is more advisable compared to using the SBFS method.

CONCLUSION

This study indicates that the average accuracy value of utilizing the SFFS method is 96.63% while utilizing the SBFS method is 82,41%. The most influential feature to the occurrence of forest fire based on SFFS is earth surface feature with a score of 97.29, while the most influential feature to the occurrence of forest fire based on SBFS is cloud coverage with a score of 88.75. All feature data used are based on weather and land coverage. Overall, there are ten features that are used in the selection process and feature exploration of which feature is related to the occurrence of a forest fire. It is concluded that the purpose of identifying the determinant parameter in a forest fire is achieved.

ACKNOWLEDGMENT

The authors would like to thank The Ministry of Research Technology and Higher Education Indonesia and Research Center Universitas Mercu Buana for supporting the research.

REFERENCES

Ali, W., Shamsuddin, S. M., & Ismail, A. S. (2012). Knowledge-Based Systems Intelligent Naïve Bayes-based approaches for Web proxy caching. *Knowledge-Based Systems*,

31, 162-175. https://doi.org/10.1016/j.knosys.2012.02.015

Alkhatib, A. A. A. (2014). A review on forest fire detection techniques. *International Journal of Distributed Sensor Networks*, 10(3), 1-12. https://doi.org/10.1155/2014/597368

Badarinath, K. V. S., Kharol, S. K., & Chand, T. R. K. (2007). Use of satellite data to study the impact of forest fires over the northeast region of India. *IEEE Geoscience and Remote Sensing Letters*, *4*(3), 485–489. https://doi.org/10.1109/LGRS.2007.896738

Biswas, S., Lasko, K. D., & Vadrevu, K. P. (2015). Fire Disturbance in Tropical Forests of Myanmar-Analysis Using MODIS Satellite Datasets. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(5), 2273–2281. https://doi.org/10.1109/JSTARS.2015.2423681

Changhui, Y., Shaohong, S., Jun, H., & Yaohua, Y. (2010). An object-based change detection approach using high-resolution remote sensing image and GIS data. In *2010 International Conference on Image Analysis and Signal Processing*, Zhejiang, China. (565–569). http://doi.org/10.1109/IASP.2010.5476052

Chuan-xu, W., Yun, L., & Zuo-yong, L. (2008). Algorithm Research of Face Image Gender Classification Based on 2-D Gabor Wavelet Transform and SVM. In *2008 International Symposium on Computer Science and Computational Technology*, Shanghai, China. (pp. 312–315). https://doi.org/10.1109/ISCSCT.2008.204

FAO. (2010). Asia-Pacific Forests and Forestry to 2020 Report of The Second Asia-Pacific Forestry Sector Outlook Study. In *Second Asia-Pacific Forestry sector Outlook Study*. Bangkok.

Fitrianah, D. (2015). Cube Density Based Spatio-Temporal Clustering Algorithm (IMSTAGRID) and Its Application in Identifying the Tuna Potential Fishing Zones in Indonesia Sea. *International Journal of Software Engineering and Its Application*.

Fitrianah, Devi, & Wasito, I. (2012). Hybrid record linkage model for integrating marine data. *Procedia Engineering*, *50*, 926–932. https://doi.org/10.1016/j.proeng.2012.10.100

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, *3*(3), 1157–1182. https://doi.org/10.1016/j.aca.2011.07.027

Hernández-Leal, P. A., González-Calvo, A., Arbelo, M., Barreto, A., & Alonso-Benito, A. (2008). Synergy of GIS and remote sensing data in forest fire danger modeling. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *1*(4), 240–247. https://doi.org/10.1109/JSTARS.2008.2009043

Joshi, S., & Nigam, B. (2011). Categorizing the document using multi class classification in data mining. In *Proceedings - 2011 International Conference on Computational Intelligence and Communication Networks*, Gwalior, India. (pp. 251–255). https://doi.org/10.1109/CICN.2011.50

Khabarov, N., Moltchanova, E., & Obersteiner, M. (2008). Valuing weather observation systems for forest fire management. *IEEE Systems Journal*, *2*(3), 349–357. https://doi.org/10.1109/JSYST.2008.925979

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kusmana, C., Istomo, Wilarso, S., Dahlan, E. N., & Onrizal. (2004). Upaya Rehabilitasi Hutan Dan Lahan Dalam Pemulihan Kualitas Lingkungan. *Seminar Nasional Lingkungan Hidup dan Kemanusiaan 2004*, Jakarta, Indonesia. (pp. 1–10).

Lellolsima, S. (2015). Kebakaran Hutan Dan Lahan Peringatan Dini BMKG dan LAPAN Kurang Diperhatikan Pemerintah. Retrieved September 11, 2018, from https://www.lapan.go.id/index.php/subblog/read/2015/2004/KEBAKARAN-HUTAN-DAN-LAHAN-Peringatan-Dini-BMKG-dan-LAPAN-Kurang-Diperhatikan-Pemerintah/677

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, *15*(11), 1119–1125. https://doi.org/10.1016/0167-8655(94)90127-9

Reddy, M. P., Prasad, B. E., Reddy, P. G., & Gupta, A. (1994). A Methodology for Integration of Heterogeneous Databases. *IEEE Transactions on Knowledge and Data Engineering*, *6*(6), 920–933. https://doi.org/10.1109/69.334882

Tanase, M., & Gitas, I. Z. (2008). An examination of the effects of spatial resolution and image analysis technique on indirect fuel mapping. In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *1*(4), 220–229. https://doi.org/10.1109/JSTARS.2009.2012475

Umamaheshwaran, R., Bijker, W., & Stein, A. (2007). Image mining for modeling of forest fires from Meteosat images. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(1), 246–253. https://doi.org/10.1109/TGRS.2006.883460

Orozco, C. V., Tonini, M., Conedera, M., & Kanveski, M. (2012). Cluster recognition in spatial-temporal sequences: The case of forest fires. *GeoInformatica*, *16*(4), 653–673. https://doi.org/10.1007/s10707-012-0161-z

Wagtendonk, J. W. Van, & Lutz, J. a. (2007). Fire Regime Attributes of Wildland Fires in Yosemite National Park, USA. *Fire Ecology,* 3(2), 34–52. http://doi.org/10.4996/fireecology.0302034

Wang, J., Wang, M., Li, P., Liu, L., Zhao, Z., Hu, X., & Wu, X. (2015). Online Feature Selection with Group Structure Analysis. *IEEE Transactions on Knowledge and Data Engineering*, *27*(11), 3029–3041. https://doi.org/10.1109/TKDE.2015.2441716

Xiaowei, L. (2014). Application of Decision Tree Classification Method Based on Information Entropy to Web Marketing. In *2014 Sixth International Conference on Measuring Technology and Mechatronics Automation*, Zhangjiajie, China. (pp. 121–127). https://doi.org/10.1109/ICMTMA.2014.34

Zhao, J., Zhang, Z., Han, S., Qu, C., Yuan, Z., & Zhang, D. (2011). SVM based forest fire detection using static and dynamic features. *Computer Science and Information Systems*, *8*(3), 821–841. https://doi.org/10.2298/CSIS101012030Z

Zhao, Y., Zhou, Z., & Xu, M. (2015). Forest Fire Smoke Video Detection Using Spatiotemporal and Dynamic Texture Features. *Journal of Electrical and Computer Engineering*, *2015*, ID 706187, 1–8. https://doi.org/10.1155/2015/706187

Zhong, Y., Zhao, J., and Zhang, L. (2014). A Hybrid Object-Oriented Conditional Random Field Classification Framework for High Spatial Resolution Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(11), 7023–7037. https://doi.org/10.1109/TGRS.2014.2306692