Printed in the Republic of Korea
ANALYTICAL SCIENCE & TECHNOLOGY
Vol. 34 No. 5, 231-239, 2021
https://doi.org/10.5806/AST.2021.34.5.231

Application and evaluation of machine-learning model for fire accelerant classification from GC-MS data of fire residue

Chihyun Park[1,★], Wooyong Park[1], Sookyung Jeon[1], Sumin Lee[1], and Joon-Bae Lee[2]

[1]*Daejeon District Office, National Forensic Service, Daejeon 34054, Korea*
[2]*Daegu District Office, National Forensic Service, Chilgok 39872, Korea*

(Received August 25, 2021; Revised September 13, 2021; Accepted September 13, 2021)

Abstract: Detection of fire accelerants from fire residues is critical to determine whether the case was arson or accidental fire. However, to develop a standardized model for determining the presence or absence of fire accelerants was not easy because of high temperature which cause disappearance or combustion of components of fire accelerants. In this study, logistic regression, random forest, and support vector machine models were trained and evaluated from a total of 728 GC-MS analysis data obtained from actual fire residues. Mean classification accuracies of the three models were 63 %, 81 %, and 84 %, respectively, and in particular, mean AU-PR values of the three models were evaluated as 0.68, 0.86, and 0.86, respectively, showing fine performances of random forest and support vector machine models.

Key words: GC-MS, machine learning, random forest, support vector machine

1. Introduction

The detection of fire accelerants in fire residues is crucial in differentiating fires caused by arson from general fires. Gasoline, kerosene, diesel, organic solvents, and candles are commonly used fire accelerants in arson.[1-3] The most widely used method for analyzing fire residues collected from the fire scene is gas chromatography-mass spectroscopy (GC-MS),[4-6] which allows the separation of each component among the numerous materials. To extract fire accelerants from residues, several extraction techniques are performed including the adsorption of low-boiling-point compounds by solid-phase micro-extraction (SPME), or extraction of high-boiling-point compounds through the solvents such as dichloro-methane or ethyl ether. The resulting full-spectrum of GC/MS data is used to verify the presence of fire accelerants, and through a series of steps illustrated as a decision tree in *Fig.* 1, various components are detected and classified.

Decision trees are used in a variety of fields as the most fundamental form of expert systems in the standardization of data interpretation with enhanced efficiency.[7-9] However, naive application of a decision tree cannot garantee reasonable trust level in the detection of fire accelerants from fire residues owing to the denaturation or loss of critical components

★ Corresponding author
Phone : +82-(0)42-866-4632 Fax : +82-(0)42-862-8074
E-mail : pch0938@korea.kr

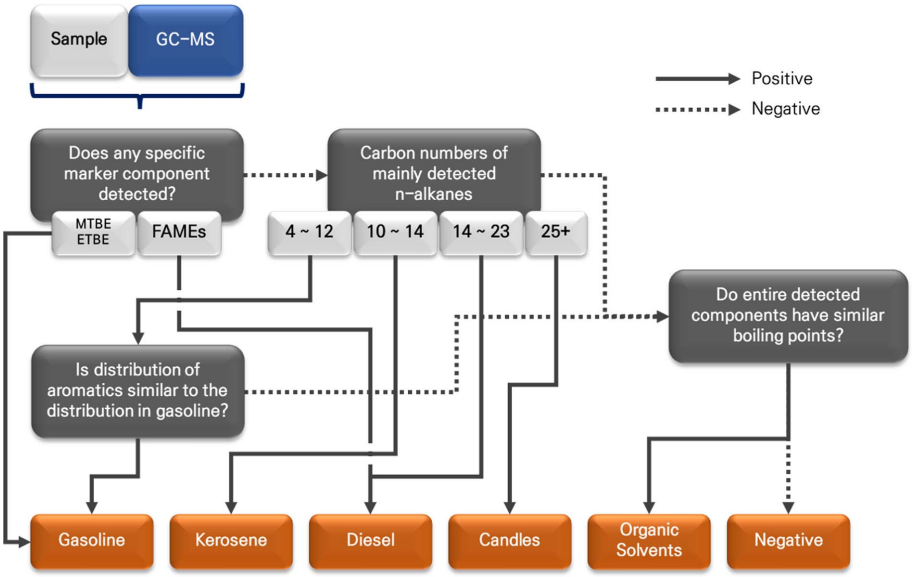

Fig. 1. Example of simple decision tree for detection of fire accelerants.

through the high temperature and combustion process in the fire scene, as well as the generation of compounds through the combustion of various petroleum-based synthetic materials such as resin, plastic, and fiber.[10] Therefore, the investigator must have a high level of expertise for accurate detection. In modern society with the trend of continuous increase in property damage due to fire, the more emphasis has been placed on fire analysis steadily, with a concurrent rise in the demand for a standard analytical model to determine the presence of fire accelerants from fire residues.

In general, a supervised learning method is used for designing a model which learns classification criteria through the labeled data among the machine-learning methods. There are several models known to exhibit high performance in classification such as the decision tree that mimics the classification criteria employed by humans, the logistic regression model with relatively simple structures and advantages in avoiding overfitting or computing resource acquisition,[11] the random forest model that assembles multiple decision trees to solve the problems of overfitting or reduced degree of freedom due to sequential categorization, which is an inherent limitation of decision trees,[12] and the support vector machine that explores the optimum hyperplane categorizing the vector data projected from a low dimension to a high dimension using various kernel functions.[13] In this study, the classification performance of the logistic regression, random forest, and support vector machine models in classifying fire accelerants was compared based on the dataset acquired from actual fire residues, and the quantitative data for marker components were also extracted from the GC-MS data.

2. Methods

2.1. GC/MS

Raw data were obtained from analyzing 728 cases on fire residues in fire accidents that occurred between 2018 and 2020 in the regions administered by the National Forensic Service Daejeon Institute. For the data analysis, GC-MS (Agilent technologies GC6890N / MS5975C, Santa Clara, CA, USA) was used, and the Supelco SPME fiber assembly carboxen/polydimethylsiloxane with film thickness(d_f) 75 μm and needle size of 23 ga (Sigma-Aldrich, USA) was used for the solid-phase microextraction of each case. For the solvent extraction, diethyl ether (Dae-jung

Table 1. Analysis conditions of GC-MS

GC-MS Condition	
Inlet	270 °C, 4.8 psi
Oven	40 °C (3 min) → (10 °C/min) → 280 °C (10 min)
Carrier gas	He 0.8 mL/min
Ion source temp.	230 °C
Column	HP-5 Capillary GC Column (30 m)
Split ratio	5:1 (SPME), 20:1 (solvent extraction)
Solvent delay	0.2 min (SPME), 2.5 min (solvent extraction)
SPME	Carboxen/polydimethylsiloxane, df 75 μm, needle size 23 ga

Chemicals, Siheung, South Korea) was used. The analytical conditions for GC-MS are presented in detail in *Table* 1.

2.2. Quantitative data interpretation and preprocessing

The GC-MS data were obtained as the mass to charge ratio (m/z) of the ions detected using GC-MS according to the retention time (Rt). To apply such basic data in supervised learning, the label and feature of each data must be defined and compiled in a database. All data were labeled as fire accelerant –, gasoline +, kerosene +, diesel +, organic solvents +, and candles +. As depicted in *Fig.* 2, a total of 42 chemical species including methyl tert-butyl ether (MTBE), ethyl tert-butyl ether (ETBE), tertiary amyl methyl ether (TAME), toluene, fatty acid methyl esters (FAMEs), C_2-Benzenes, C_3-Benzenes, C_4-Benzenes, naphthalene, and n-alkanes (carbon number 6-26), which are widely used in the detection of petroleum-based fossil fuels, were selected as the markers in the quantitative criteria of the base peak intensity of each compound based on the GC-MS data. As the absolute quantity for each detected compound significantly varied according to the analytical conditions and states of samples, the results were standardized as *Px*, which is calculated by dividing the base peak intensity of specific component *x* by sum of the base peak intensity of every marker detected in a given sample.

2.3. Cross-validation and dataset preprocessing for learning

The overall design of classification models is illustrated in *Fig.* 3. To compare the performance across the logistic regression, random forest, and support vector machine models under identical conditions, the overall structure was maintained constant except for the type of the classification model. In data preprocessing, all input values scaled as $\ln(1+Px)$ to prevent the underestimation of the presence of a marker when the fraction indicated a trace amount. To verify the

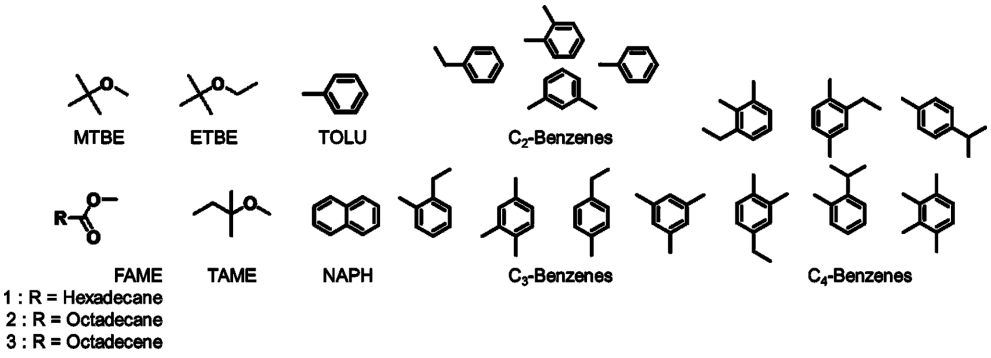

Fig. 2. Representative marker components for hydrocarbon-based fire accelerants.

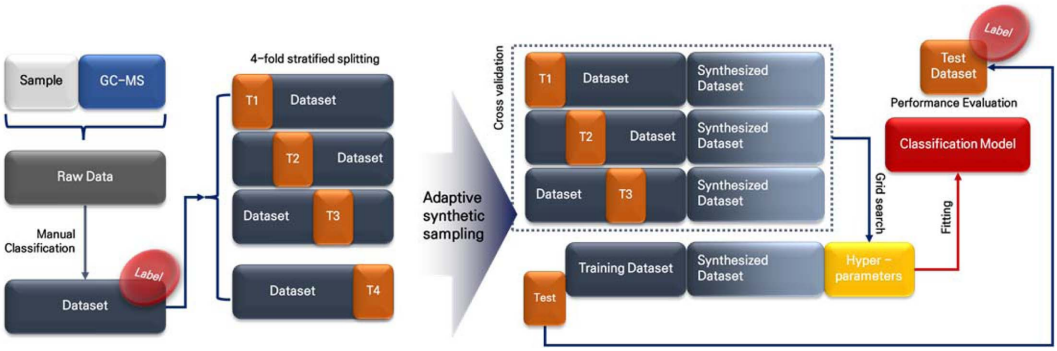

Fig. 3. Schematic diagram of data preprocessing, learning and evaluation process for classification models.

validity of each model and explore the hyperparameters in model optimization through the datasets in this study, three of the four stratified data folds were used in the hyperparameter search and cross-validation, while the remaining data fold was used in the training and test steps, with each dataset consisting of 80 % training sets and 20 % test sets.

The dataset used in this study have a severe data imbalance across labels; there were only 8 cases of kerosene + and 17 cases of candles +, whereas there were 351 cases of fire accelerant – and 207 cases of gasoline +. Such an imbalance in input data labeling is known to cause an unintended bias based on the input data quantity in the model training and test sets to ultimately degrade the model performance.[14] To solve this problem prior to training, upsampling was applied to set an identical data quantity in each label of the training set during cross-validation and model learning. Adaptive synthetic sampling (ADASYN) was used to achieve a high level of performance upon the upsampling of continuous variables.[15]

2.4. Classification model hyperparameter search and performance evaluation

The representative hyperparameter in logistic regression is the C value, which is defined as the reciprocal of the regularization parameter λ, where lower C values indicate more conservative decisions in the decision boundary and higher C values indicate a higher fitting to the dataset with a higher risk of overfitting. The mean classification accuracy and standard deviation according to the change in the C value of the cross-validation dataset are presented in *Table* 2. The use of L2 regularization, 10^{-4} tolerance, and Limited Memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) method as the search algorithm led to an increase in the accuracy of the model between 10^{-5} and 10^{5} C values, after which the accuracy decreased owing to overfitting. The standard deviation for the higher mean accuracy at 0.639 was 0.042.

Table 2. Mean accuracy of logistic regression models at varying C values

C	Mean accuracy ± SD	C	Mean accuracy ± SD
10^{-4}	0.110 ± 0.023	10^{2}	0.619 ± 0.029
10^{-3}	0.123 ± 0.005	10^{3}	0.632 ± 0.035
10^{-2}	0.167 ± 0.023	10^{4}	0.637 ± 0.043
10^{-1}	0.322 ± 0.034	10^{5}	**0.639 ± 0.042**
10^{0}	0.494 ± 0.020	10^{6}	0.637 ± 0.039
10^{1}	0.584 ± 0.038	10^{7}	0.632 ± 0.044

In the random forest model, an increase in the number of estimators causes an increase in performance but a simultaneous proportional increase in the required resources for the calculation, while an increase in maximum depth causes an increase in accuracy but a simultaneous increase in the probability of overfitting. As presented in *Table* 3, the evaluation of the cross-validation dataset regarding the number of decision trees and maximum depth of the optimized random forest model applying the Gini coefficient and bootstrap shows that the mean classification accuracy was 0.793 with a standard deviations of 0.026 for 10^{3} decision

Table 3. Mean accuracy of random forest models at varing hyperparameters

Random forest		Number of estimators				
		10^0	10^1	10^2	10^3	10^4
Max depth	10	0.553 ± 0.005	0.736 ± 0.027	0.790 ± 0.023	**0.793 ± 0.026**	0.780 ± 0.034
	20	0.552 ± 0.058	0.732 ± 0.021	0.780 ± 0.045	0.786 ± 0.030	0.782 ± 0.027
	30	0.552 ± 0.039	0.720 ± 0.052	0.768 ± 0.030	0.782 ± 0.026	0.782 ± 0.032
	40	0.534 ± 0.007	0.731 ± 0.012	0.791 ± 0.027	0.784 ± 0.029	0.784 ± 0.029

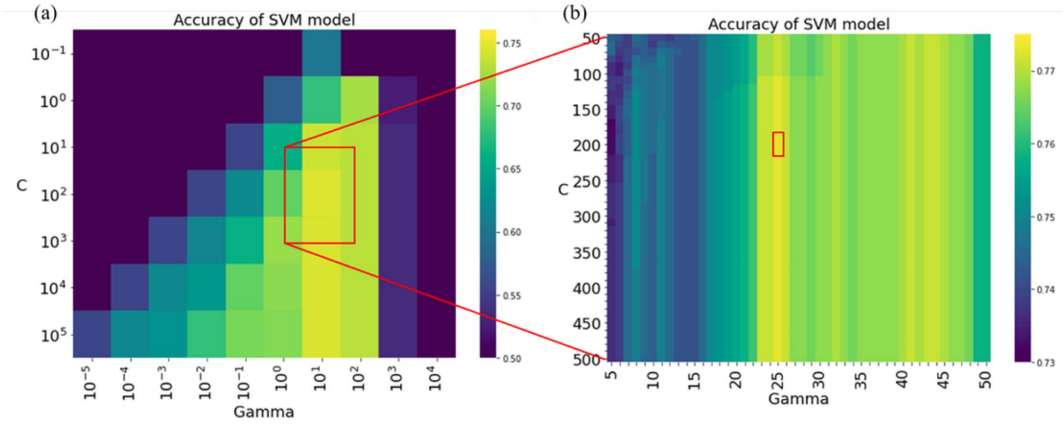

Fig. 4. Mean accuracy of support vector machine models at varying its hyperparameters C and γ, in (a) log-scaled broad range and (b) more specific range.

trees and a maximum depth of 10 layers.

The support vector machine using the radial basis function (RBF) as a kernel is generally known to exhibit a high classification performance compared with those of the logistic regression or random forest models, while it is sensitive to the C value and the kernel variable γ. The C value-related trend in the model is the same as that in the logistic regression model. If γ is too small, it is difficult to obtain an adequate classification performance as the hyperplane searched by the model approaches a linear form. If γ is too large, the hyperplane takes a form that indicates an increased risk of overfitting the dataset. As depicted in *Fig.* 4, the support vector machine model at 10^{-3} tolerance for the cross-validation dataset yielded the highest performance for the hyperparameter C = 200 and γ = 25, where the mean classification accuracy was 0.773 with a standard deviation of 0.023.

To test the trained models based on optimized hyperparameters, the classification accuracy of each model was expressed in confusion matrices. In addition, the area under receiver operation characteristics (AU-ROC) and the area under precision-recall (AU-PR) were calculated for each model as general indicators of the model classification performance.

3. Results and Discussion

3.1. Quantitative data interpretation

Fig. 5 depicts the mean *Px* of each marker component as calculated from the cases which were categorized as fire accelerant – , gasoline +, kerosene +, diesel +, organic solvents +, and candles +, respectively. As shown, various marker components were detected even in the absence of fire accelerants due to the components formed during a fire. MTBE, ETBE and TAME were detected only in the case of gasoline +, in contrast to the cases of other petroleum-based fuels, with an abundance of aromatic compounds in comparison to n-alkanes. The detection of n-alkanes of carbon numbers 9-16 and 17-21 was characteristic of

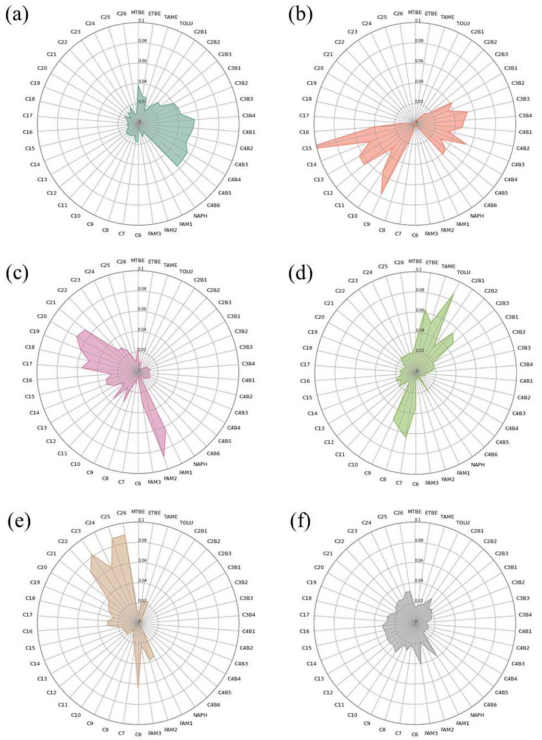

Fig. 5. Mean base peak intensity ratio of each marker component from GC-MS data of fire residues, which were concluded as (a) gasoline, (b) kerosene, (c) diesel, (d) organic solvents, (e) candles, and (f) no fire accelerants were involved.

kerosene + and diesel + respectively, and in particular, in the case of diesel +, the FAMEs that are not commonly found in other fire accelerants were detected. In the case of organic solvents +, low-boiling-point compounds rather than a specific set of chemical species were mainly detected, which may be attributed to the process-dependent distribution of components in petroleum-based organic solvents whose purification involves thermal distillation. In the case of candles +, a notable characteristic was a high detection rate of n-alkanes of carbon numbers above 20.

3.2. Model performance evaluation

Table 4 presents the performance evaluation results of the logistic regression, random forest, and support vector machine models after the training using the training set. To evaluate the performance of the model, the ratio of positive identifications among actual positive instances (precision), the ratio of actual positives among positive identified instances (recall), and the overall accuracy of data classification were examined. The lowest classification accuracy was 63 % for the logistic regression model, followed by the random forest (81 %) and support vector machine (84 %) models. To determine the detailed error trends, the confusion matrices for each model were expressed as in *Fig.* 6. All three models exhibited the lowest recall for kerosene with the smallest number of cases. A false-positive trend was observed for the logistic regression model, where approximately 46 % of the fire accelerant – data was identified as positive, with less than 70 % recall across all classifications except gasoline and candles, indicating a relatively low performance. On the contrary, the random forest model exhibited a considerably high recall at 82 % for the fire accelerant – data, with the rate of recall for gasoline + and diesel

Table 4. Performance evaluation of classification models

	Logistic regression		Random forest		Support vector machine	
	Precision	Recall	Precision	Recall	Precision	Recall
Negative	0.84	0.55	0.88	0.82	0.90	0.92
Gasoline	0.76	0.79	0.77	0.88	0.80	0.87
Kerosene	0.33	0.50	1.00	0.50	0.50	0.50
Diesel	0.47	0.69	0.69	0.69	0.62	0.62
Solvent	0.42	0.57	0.74	0.74	0.88	0.65
Candle	0.17	0.75	0.67	0.50	0.75	0.75
Accuracy	0.63		0.81		0.84	
AU-ROC	0.90		0.96		0.96	
AU-PR	0.68		0.86		0.86	

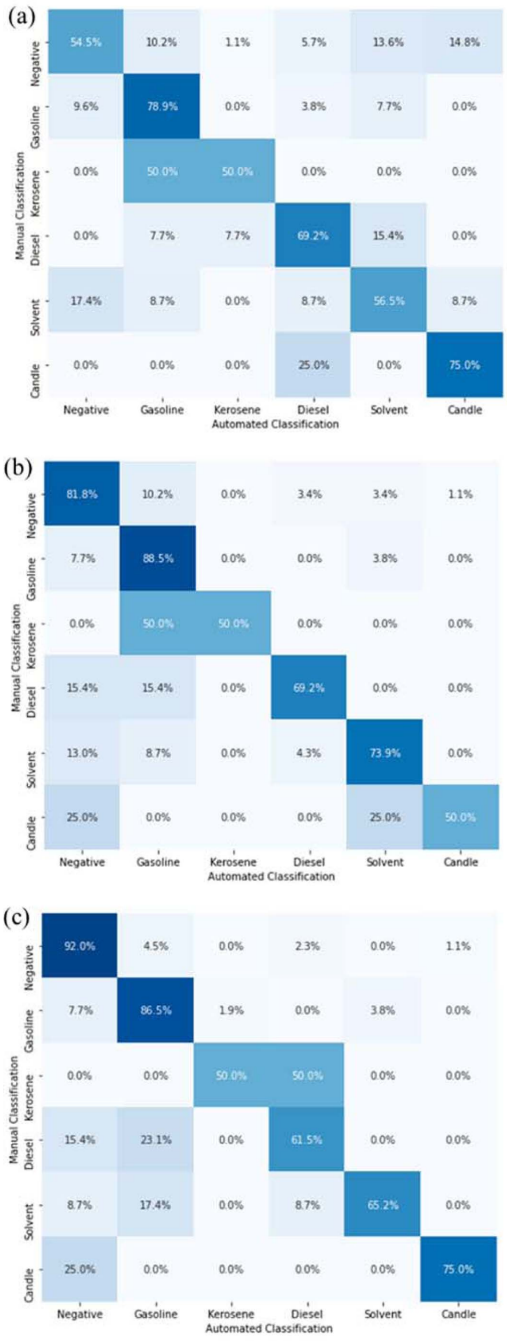

Fig. 6. Confusion matrices of (a) the logistic regression, (b) the random forest, and (c) the support vector machine models which are trained to classify fire accelerants.

+ being the highest among the three models at 89 % and 74 %, respectively. Lastly, the rate of recall for the fire accelerant – data in the support vector machine model was 92 %, implying the lowest risk of false positives. Both the random forest and support vector machine models were found to exhibit a relative vulnerability to the error of identifying the data of candles + as fire accelerant – (25 % in both models) or diesel + as gasoline + (15 % and 23 %, respectively).

Fig. 7 depicts the ROC curves and PR curves across all classifications in each model, and the area under (AU) of the curves to serve as the performance indicator of positive-negative determination was calculated and is presented in *Table* 4. The models are considered to be reliable if the AU-ROC approximates to 1 as the ROC approaches the top left corner, indicating that the model clearly distinguishes positive or negative. The AU-ROC was 0.90, 0.96, and 0.96, respectively, for the logistic regression, random forest, and support vector machine models. Though all three models show valid discriminative power, the random forest and support vector machine models were shown to more accurately identify each instance than logistic regression model. In general, the validity of a classification model is adequately determined based on AU-ROC; however, due to the severe imbalance of the dataset used in this study, AU-PR was additionally assessed for meaningful performance comparison. The level of confidence in a positive determination by the model increases as the average PR curve approaches the top, which take the mean proportion of positives across all classification data as the baseline. The curve above the baseline verifies the validity of the model, and as the AU-PR approaches 1, the model can be considered to indicate a higher performance. The AU-PR of the logistic regression, random forest, and support vector machine models was 0.68, 0.86, and 0.86, respectively, all within a valid range; however, the performance was substantially lower for the logistic regression model than that of the other two models, coinciding with the trend observed in the confusion matrices.

4. Conclusions

In this study, for developing the model for automated

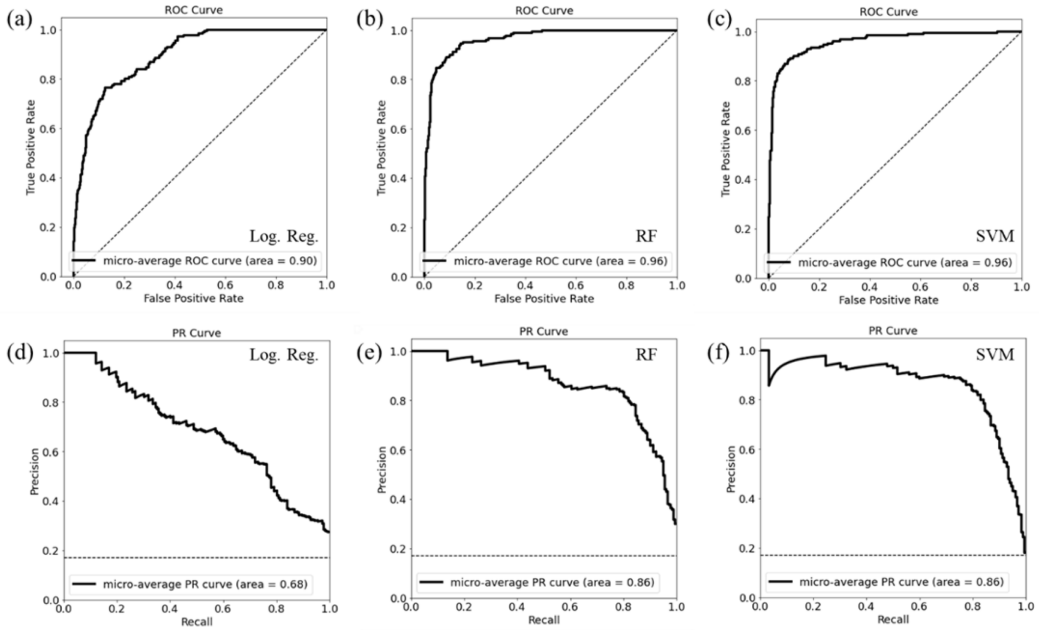

Fig. 7. ROC curves (a, b, c) and PR curves (d, e, f) of the logistic regression (a, d), the random forest (b, c), and the support vector machine (c, f) models.

fire accelerant classification, datasets were created for the indicators of fire accelerants using the GC-MS data from 728 cases on fire residue. Cross-validation and hyperparameter optimization were performed, while the logistic regression, random forest, and support vector machine models were assessed after the training under optimum conditions. The classification accuracy was the highest (84 %) for support vector machine, followed by the random forest (81 %) and the logistic regression (63 %) models. Considering the ground truth of the data, the three models yielded sufficiently high performance. Notably, AU-PR was measured as 0.86, 0.86, and 0.68 respectively, ensuring a high level of performance of support vector machine and random forest models.

Main limitation of this study comes from dataset imbalance across labels, which caused inevitable vulnerability of the models. In particular, the classification accuracy was low for kerosene with a small dataset, although the effect of the dataset imbalance was minimized through ADASYN for data preprocessing step and measuring AU-PR for performance evaluation. Based on the results, the validity was verified for the random forest and support vector machine models through the GC-MS data. Further studies should pursue additional assessments and cross-validations toward the enhanced performance and validity of the models via a continuous collection of the GC-MS data on fire residues for model learning and validation, as well as the introduction of the validation set using isolated datasets.

Acknowledgements

This work was supported by National Forensic Service (NFS2021CHE02), Ministry of the Interior and Safety, Republic of Korea.

References

1. R. M. Smith, *Anal. Chem.*, **54**(13), 1399A-1409A (1982).
2. A. Hamins, M. Bundy and S. E. Dillon, *J. Fire Prot. Eng.*, **15**(4), 265-285 (2005).
3. A. D. Pert, M. G. Baron and J. W. Birkett, *J. Forensic Sci.*, **51**(5), 1033-1049 (2006).
4. S. T. Teng, A. D. Williams and K. Urdal, *J. High.*

Resolut. Chromatogr., **17**(6), 469-475 (1994).
5. R. O. Keto, *J. Forensic Sci.*, **40**(3), 412-423 (1995).
6. C. H. Wu, C. L. Chen, C. T. Huang, M. R. Lee and C. M. Huang, *Anal. Lett.*, **37**(7), 1373-1384 (2004).
7. J. N. Eisenberg and T. E. McKone, *Environ. Sci. Technol.*, **32**(21), 3396-3404 (1998).
8. E. Akkaş, L. Akin, H. E. Çubukçu and H. Artuner, *Comput. Geosci.*, **80**, 38-48 (2015).
9. M. G. Yıldız, T. Davran-Candan, M. E. Günay and R. Yıldırım, *J. CO2 Util.*, **31**, 27-42 (2019).
10. D. C. Mann, *J. Forensic Sci.*, **32**(3), 616-628 (1987).
11. J. S. Cramer, The Origins of Logistic Regression, http://dx.doi.org/10.2139/ssrn.360300, Assessed 25 Jan 2003.
12. L. Breiman, *M. Lear.*, **45**(1), 5-32 (2001).
13. C. Cortes and V. Vapnik, *M. Lear.*, **20**(3), 273-297 (1995).
14. A. Luque, A. Carrasco, A. Martín and A. de las Heras, *Pattern Recognit.*, **91**, 216-231 (2019).
15. H. He, Y. Bai, E. A. Garcia and S. Li, 'ADASYN: Adaptive synthetic sampling approach for imbalanced learning', IEEE, 2008.

Authors' Position

Chihyun Park : Researcher
Wooyong Park : Research Officier
Soo-Kyung Jeon : Researcher
Soo-min Lee : Research assistant
Joon-Bae Lee : Research Officier