Smoke Detection with Ensemble Modeling

Pongsakorn Teerarassamee*, Ratiporn Chanklan, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—This research aims at investigating performance of the ensemble learning method. The ensemble learning brings together various weak learners to create strong learners. Based on this ensemble learning idea, we develop a model for an efficient smoke detection tool. The three schemes of ensemble learning are investigated including bagging, boosting, and stacking. The bagging ensemble algorithm studied in this research is Random Forest and the boosting algorithm is AdaBoost. The stacking ensemble adopts three algorithms, that are Random Forest, AdaBoost, and Logistic Regression. The other learning algorithms adopted for performance comparison include Support Vector Machine, Naïve Bayes, and Decision Tree. The smoke detection data contain 62,630 records and 15 features. The dataset has been separated into training set and test set with a ratio of 75:25. The experimental results reveal that AdaBoost outperforms other learning algorithms when applied to the specific smoke detection application domain.**

***Index Terms*—Smoke detection, ensemble learning, weak learner, bagging, boosting, stacking**

I. Introduction

The importance of technology in modern life is on the rise. Whether it is a convenience issue or even storage, it was designed to record diverse information in digital files when technology started to play a role in helping to preserve information from the past that had to be transcribed, written, or recorded on paper, even an information storage database system. When having convenient access to information, data analysis is now quicker and easier than it was in the past. One of the common approaches is to choose to examine data using computer techniques. By building a model for prediction, data classification techniques are also mentioned. Such approaches outcomes are categorized as one of the machine learning disciplines, which is a common practice today.

Computerized data analysis has span over various domains whether it is a company that makes use of customer predictions for purchases, or the branch of medicine that forecasts a patient's illness [1]. Furthermore, specialized data can also be subjected to classification procedures. In this study, we use a smoke detection data classification strategy to develop a prediction model with ensemble technique. Our purpose is that if data on smoke detection can be used to develop a reliable model, better forecast outcomes will enable the prevention or mitigation of fire damage. Since the objective is to create an AI-based smoke detector gadget, IOT devices are used to collect data [5]. To guarantee a good dataset for training, a variety of habitats and fire sources must be sampled. This study also uses these data in the experiment. According to research on smoke detection data, numerous operations include image processing to identify forest fires [6] are applied.

In this research, the model has been built using the ensemble learning technique, and then evaluate its performance by contrasting it with other modeling schemes. Random Forest, AdaBoost, and Logistic Regression are used in ensemble learning for comparison. The other three learning schemes applied for the comparison purpose including the C-SVC algorithm (the type of support vector machine), the Naïve Bayes algorithm which is the probabilistic classifier, and Decision Tree algorithm to be applied as a representative of weak learner type. Comparison by type is summarized in the Table I. Model performance is to be compared based on the four-measurement metrics: accuracy, precision, recall, and F1 score. We investigate modeling performance using ensemble learning, SVM, amd Naïve Bayes because they are often used in many research such as detecting fake hotel reviews [3] or wind speed classification [4].

For the purpose of modeling the data and testing the model performance, the data are split into two subsets: training data and test data. The train-test evaluation method had been widely used in several studies and those researchers divided data into many ratios such as 70:30, 75:25, and 80:20 [2]. Based on our large data size, this study will make the train-test data ratio to be 75:25.

TABLE I: Overview of Applied Algorithms

Type	Algorithm
Ensemble Learning - Bagging	Random Forest
Ensemble Learning - Boosting	AdaBoost
Ensemble Learning - Stacking	Random Forest, AdaBoost, Logistic Regression
Weak Learner	Decision Tree
Support Vector Machines	C-Support Vector Classification
Probabilistic Classifiers	Naïve Bayes

II. Theory

A. Ensemble Learning

Ensemble learning is one of the machine learning strategies that focuses on enhancing the performance of the model by including many weak learners in order to get better outcomes. The strong learner model is another name for it. Three methods of ensemble learning [7] can be categorized: bagging, boosting, and stacking.

Manuscript received January 20, 2023; revised February 20, 2023; accepted March 24, 2023; published April 15, 2023.

P. Teerarassamee, K. Kerdprasop and N. Kerdprasop are with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand.

R. Chanklan is with the Faculty of Engineering and Technology, School of Computer Engineering, Rajamangala University of Technology Isan, Nakhon Ratchasima, Thailand.

*Correspondence: psk.trrsm@gmail.com (P.T.)

DOI: 10.7763/IJMO.2023.V13.823

B. Bagging

Bootstrap aggregating, often known as bagging, is a method that enables the segmentation of unstable processes [8] to produce ensemble learning models that are resilient to data volatility. By utilizing weak learners to model each component in numerous portions, this procedure can run at the same time in a parallel manner. The final result can be obtained from the average of the numerical data. Random Forest is a well-known and widely used algorithm [9]. Additionally, Preimage learning employs this technique [10].

C. Boosting

Boosting is the ensemble learning method that combines several learners and processes sequentially. At first, all the data were extracted to build a flimsy learner model. Then, continue creating a weak learner model by adding the incorrect component to improve the data. The process continues until the outcome bias is decreased [11]. AdaBoost is a well-known algorithm in this boosting ensemble category and it is successfully applied in many application domains [12]. Ensemble learning is built on the concept of the weak learner using a decision tree because the vulnerability of the tree can be thoroughly controlled. The decision tree algorithm is typically selected, or even the way the tree is arranged with only one node. Tree growing is based on the choices using only one factor providing decision-related information in a binary classification. Decision tree stump refers to this classification [13].

D. Stacking

Stacking technique includes several models and algorithms. This extends beyond bagging and boosting [14]. The stacking scheme can combine models from several schemes such as AdaBoost, Random Forest, and Logistic regression, all at once.

E. Evaluation

Common measurement for assessing classification models is accuracy. Accuracy is the percentage of correct predictions per overall predictions and it can be computed as in Eq. (1) when the meaning of each acronym is summarized in Table II.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

When evaluating the effectiveness of a machine learning model, precision is a critical feature to consider. It is described as the ratio of true positive results to all positive forecasts, including accurate markers and false positives, as in Eq. (2).

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

Another important indicator of an efficient machine learning model is recall. It allows to compare the number of accurate items found to the number of actual items. Recall has the computation as shown in (3).

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

A powerful metric for evaluating overall performance of a model is the F1 score, or the F-measure. Precision and recall are two metrics that are combined in F1 score and its formula is as in Eq. (4).

$$F1\,Score = \frac{2TP}{2TP+FP+FN} \tag{4}$$

By evaluating the model's output with the actual result from the test data, efficiency calculations may be made. This technique is categorized as a confusion matrix [15] and it is used to contrast performance indicators like those in Table II.

TABLE II: CONFUSION MATRIX

		Predict	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

where

TP is the amount of data that the model predicts to be positive and the actual data are positive,

TN is the number of data that the model predicts to be negative, and the actual data are negative,

FP is the number of data that the model predicts to be positive, but the actual data are negative,

FN is the number of data that the model predicts to be negative, but the actual data are positive.

III. SMOKE DETECTION DATASET

Smoke detectors provide statistics on the detection of smoke. One of the things that can save lives is a smoke detector. For instance, from 1982 to 2012, the number of fire casualties in France decreased by more than 48%, while from 1982 to 2013, the number decreased by 56% in the UK. The majority of these are connected to smoke alarms and stricter fire safety standards. Smoke detectors are installed in 96% of American households, or around 20% of all homes [16]. Smoke detectors do not work when they are off. According to predictions, the number of US residential fire deaths might be cut by 36% if every home had a functional smoke detector, saving up to 1100 lives annually [17]. False fire alarms started to be a concern [18]. False fire alarms kept popping up more frequently. It poses a significant issue for firemen. Smoke detection dataset contains 62,630 records and 15 features. The description of data is summarized in the Table III.

TABLE III: SMOKE DETECTION DATASET

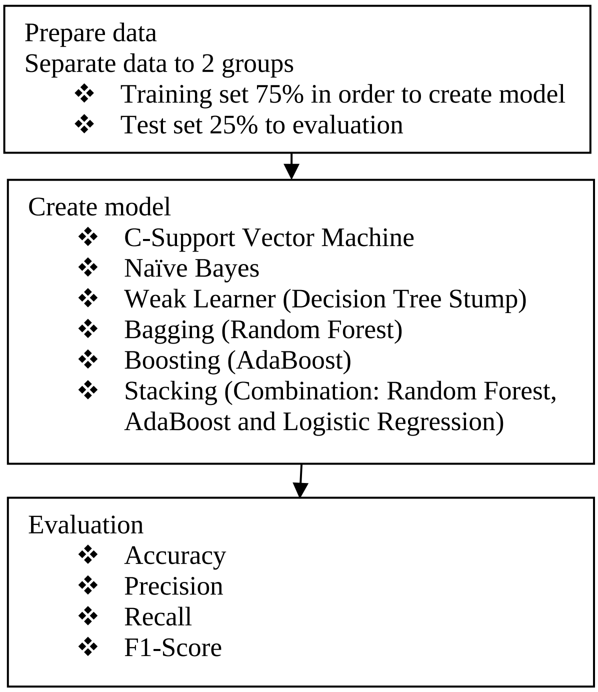
Feature Name	Data Type
Timestamp (UTC)	Time
Air temperature	Numeric
Air humidity	Numeric
Total volatile organic compounds	Numeric
Equivalent CO_2 concentration	Numeric
Coarse molecular hydrogen (generated H_2)	Numeric
Raw ethanol	Numeric
Air pressure	Numeric

Particle size < 1.0 µ m (PM 1.0)	Numeric
Particle size < 2.5 µ m (PM 2.5)	Numeric
Particle concentration < 0.5 µ m (NC0.5)	Numeric
0.5 µm < particle concentration < 1.0 µm (NC1.0)	Numeric
1.0 µm < particle concentration < 2.5 µm (NC2.5)	Numeric
Sample counter (CNT)	Numeric
Fire alarm	Label (0=non alarm, 1=alarm)

IV. Conceptual Design

The study of ensemble learning is the main topic of this research. The research procedure can be broken down into three stages: data preparation, which involves splitting the data into two subsets for training and testing; model creation; and model evaluation. The conceptual design is depicted in Fig. 1.

This study uses 6 algorithms (3 ensemble learning schemes and other 3 learning algorithms) to develop models for smoke detection. The bagging technique uses Random Forest algorithm, and boosting technique uses AdaBoost algorithm. Stacking technique combines the algorithms AdaBoost, Random Forest, and Logistic Regression.



Prepare data
Separate data to 2 groups

- Training set 75% in order to create model
- Test set 25% to evaluation

Create model

- C-Support Vector Machine
- Naïve Bayes
- Weak Learner (Decision Tree Stump)
- Bagging (Random Forest)
- Boosting (AdaBoost)
- Stacking (Combination: Random Forest, AdaBoost and Logistic Regression)

Evaluation

- Accuracy
- Precision
- Recall
- F1-Score

Fig. 1. Conceptual design framework.

Additionally, we use two other algorithms: C-Support Vector Machines and Naive Bayes as well as one weak learner (Decision Tree Stump) for performance comparison. Model is to be assessed through comparison of performance with the accuracy, F1 score, precision, and recall gauges.

V. Result

The training data have been used to build a model after the smoke detection data was split into 25% to be testing data and 75% to be training data. Then, use the testing data to evaluate the performance of models using the various measurements. Results are shown in Table IV.

Consider from the accuracy performance, the best model is the one created with the algorithm AdaBoost (accuracy = 98.85%), which is the algorithm in the category of boosting ensemble. In terms of precision, the model with the best performance in this aspect is the model created by a combination of the three algorithms using a stacking ensemble scheme. Stacking technique combines three base algorithms that are Random Forest, AdaBoost and Logistic Regression. This learning scheme obtains the precision at 100%, which is as high as the C-Support Vector Classification. When comparing recall value, the Weak Learner (Decision Tree) has the highest value (recall=100%). For the overall performance evaluation using F1 score, the AdaBoost shows the highest F1 score at 99.19%.

TABLE IV. Model Performance Measurement Results

Algorithm	Evaluation			
	Accuracy	Precision	Recall	F1 Score
C-SVC	71.13%	**100.00%**	71.13%	83.13%
Naive Bayes	83.75%	97.66%	82.65%	89.53%
Weak Learner	90.46%	86.59%	**100.00%**	92.81%
Bagging	88.82%	98.82%	87.17%	92.63%
Boosting	**98.85%**	98.63%	99.75%	**99.19%**
Stacking	71.13%	**100.00%**	71.13%	83.13%

VI. Conclusion

The purpose of this research is to study the model building to predict the smoke detection efficiency by evaluating the performance based on the four measurement metrics: accuracy, precision, recall, and F1 score. The models are created from several learning schemes including ensemble with bagging, ensemble with boosting, ensemble with stacking, weak leaner using decision tree stump, probabilistic learner using naive Bayes, and support vector machine. From the experimental results boosting ensemble technique with the AdaBoost algorithm yields the best performance in terms of F1 score. Although the accuracy and recall of the boosting technique is not the highest value, but the F value, which is a measure of performance obtained by averaging accuracy and the commemorative value gives the highest value. Therefore, it can be concluded that for the smoke detection data, a boosting technique can be used to create the predictive model.

For limitation about this study, the best result claim only the smoke detection dataset.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Pongsakorn Teerarassamee is the main contributor in designing research process, experimentation, and paper submission. Ratiporn Chanklan, Kittisak Kerdprasop is responsible for experimentation and manuscript preparation. Nittaya Kerdprasop helps revising the manuscript. The fourth author helps organizing the structure of the paper and proving the experimental results.

REFERENCES

[1] J. Jiang, H. Pan, M. Li, B. Qian, X. Lin, and S. Fan, "Predictive model for the 5-year survival status of osteosarcoma patients based on the SEER database and XGBoost algorithm," *Scientific Report*, vol.11, no. 5542, 2021.

[2] A. A. Smadi, A. Mehmood, A. Abugabah, E. Almekhlafi, and A. H. Al-smadi, "Deep convolutional neural network-based system for fish classification," *International Journal of Electrical and Computer Engineering*, vol. 12, no.2, pp. 2026-2039, 2022.

[3] M. Y. Chuttur and R. Bissonath, "A comparison of AdaBoost and SVC for fake hotel reviews detection," in *Proc. 2022 3rd International Conference on Computation, Automation and Knowledge Management (ICCAKM*), Dubai, United Arab Emirates, 2022, pp. 1-6.

[4] P. SangitaB and S. R. Deshmukh, "Use of support vector machine, decision tree and naive Bayesian techniques for wind speed classification," in *Proc. 2011 International Conference on Power and Energy Systems*, Chennai, India, 2011, pp. 1-8.

[5] S. Blattmann, *Real-time Smoke Detection with AI-based Sensor Fusion*, August, 2022.

[6] J. Zhan, Y. Hu, G. Zhou, Y. Wang, W. Cai, and L. Li, "A high-precision forest fire smoke detection approach based on ARGNet," *Computers and Electronics in Agriculture*, vol. 196, 2022.

[7] Z. Zhou, *Ensemble Methods: Foundations and Algorithms*, Chapman and Hall/CRC, 2012.

[8] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[10] A. Shinde, A. Sahu, D. Apley, and G. Runger, "Preimages for variation patterns from kernel PCA and bagging," *IIE Transactions*, vol. 46, no. 5, pp. 429-456, 2014.

[11] L. Breiman, "Arcing classifier (with discussion and a rejoinder by the author)," *The Annals of Statistics*, vol. 26, no. 3, pp. 801-849, 1998.

[12] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp.119–139, 1997.

[13] K. Max and K. Johnson, *Applied Predictive Modeling*, Springer, 2018.

[14] A. Gupta, V. Jain, and A. Singh, "Stacking ensemble-based intelligent machine learning model for predicting post-COVID-19 complications," *New Generation Computing*, vol. 40, no. 4, pp.987–1007, 2022.

[15] D.M.W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp.37–63, 2011.

[16] Fire Safety Statistics. Modern Building Alliance. [Online]. Available: https://www.modernbuildingalliance.eu/fire- safety-statistics

[17] Home Office Statistics Highlight Fire False Alarms Remain an Issue. BAFE Fire Safety Register — British Approvals for Fire Excellence, [Online]. Available: https://www.bafe.org.uk/bafe-news/ home-office-statistics-highlight-fire-false-alarms-remain-an-issue

[18] Smoke Alarm Research, National Institute of Standards and Technology. [Online]. Available: https://www.nist.gov/ el/smoke-alarm-research