Optimal Feature Selection Algorithm (FSA) for Electronic Nose Signal

Rizqy Ahsana Putri
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya 60111, Indonesia
6025231069@student.its.ac.id

Shoffi Izza Sabilla
Department of Medical Technology
Institut Teknologi Sepuluh Nopember
Surabaya 60111, Indonesia
shoffi@its.ac.id

Riyanarto Sarno
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya 60111, Indonesia
riyanarto@if.its.ac.id

2023 1st International Conference on Advanced Engineering and Technologies (ICONNIC) | 979-8-3503-0648-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICONNIC59854.2023.10467988

***Abstract*—In E-Nose data, redundancy and overlapping are frequent issues. This is due to the fact that certain sensors have the same gas target as other sensor gas targets.. The goal of this research is to classify meat odor E-Nose data into four classes, including chicken, beef, lamb, and pork. 276 meat odor data in total—108 chicken, 61 beef, 27 lamb, and 80 pork—were used in this research. Additionally, this research aims to identify the most effective Feature Selection Algorithm (FSA) to address issues with redundancy and overlapping data. In this study, FSA techniques such as Pearson Correlation, ANOVA F-value, Chi-Square, and MIFS were applied. SVM, K-NN, and Random Forest are the implementations of classification algorithms. According to the research's findings, SVM-MIFS (k = 10) is the best combination method for classifying meat odor data, achieving an accuracy rate of 98.91%.**

Keywords— Classification, Electronic Nose, Feature Selection Algorithm, Meats, Redundancy

I. Introduction

Feature Selection Algorithm (FSA) is a method used in data analysis and machine learning to identify relevant and important subsets of features (variables or attributes) from a dataset [1], [2]. FSA can also be used to overcome redundancy [3]. The existence of redundant data in a dataset is a problem that is often encountered in data analysis research. Redundant data can cause inaccurate classification results and reduce the accuracy of the model [4], [5]. Therefore, the application of FSA is an important step in a classification process or other data processing process. There are various types of FSA, e.g. Mutual Information Feature Selection (MIFS), F-value analysis of variance (ANOVA), chi-square, Monte Carlo feature selection (MCFS), Pearson Correlation, and Recursive Feature Elimination (RFE) [6], [7]. Many studies have used FSA to reduce the risk of redundancy in their datasets. In their research about text classification, Yujia Zhai et al. used two FSA methods, Chi-square and Information Gain which got the best accuracy of 96.74% [8]. Inzamam Mashood Nasir et al. use person correlation to eliminate redundant features in the document classification process [9]. Pan Huang et al. also used one of the FSA methods, i.e. ANOVA F-value to evaluate redundancy in their research about cancer classification [10]. Besides those researches, FSA is also massively applied in processing Electronic-Nose (E-Nose) signal data.

An Electronic nose is a device or system designed to detect and recognize various odors using electronic sensors that were inspired by the human olfactory system's ability to detect and distinguish between different odors. The electronic nose combines sensor technology and data processing to identify and classify different odor patterns. E-Nose generally utilizes gas sensors as its main component. The gas sensor is trying to arrest and then convert the odors into an electrical signal. In the past few years, E-Nose has been widely used to deal with various problems, such as for analyzing the pork content in beef [11], classifying tea [12], classifying perfume [13], classifying chili [14], measuring human blood sugar levels [15], monitoring the quality of cooking oil [16], etc.

Sometimes, sensors have a sensitivity to the same gases as other sensors. This is what eventually led to overlapping data and causes the sensor combination to not be optimal [17]. FSA can be a solution to determine optimal sensor combinations [18]. Many researchers about E-Nose have implemented FSA. Mhd Arief hasan, et al. implemented Pearson Correlation and Chi-square for feature selection in their research about the classification of the sweetness of pineapple using Electronic Nose [19]. Hannaneh Mahdavi et al. also implemented four FSA methods in their research, i.e. Chi-Squared, mutual information regression, Recursive Feature Elimination (RFE), and tree-based feature selection [20].

Several supervised learning methods are also commonly used in E-Nose research. Research on the evaluation of tea quality was conducted by Min Xu et al. using SVM and Logistic Regression. The research obtained very good performance results, even reaching 100% accuracy [21]. Li-Ying Chen et al. also implemented K-NN and SVM in classifying bananas based on their ripeness by utilizing the E-Nose and camera system. The results of reaching 100% accuracy using both K-NN and SVM show that the two methods have a very good ability to perform classification [22].

The purpose of this research was to classify various types of meat into 4 classes based on E-Nose data. Its major contribution was determining the most effective way to choose features in order to lessen data redundancy and produce higher accuracy. There are several sections in this paper: Section 2 describes the details of the methods used in this research. Section 3 describes the research results. The last section contains the conclusions of this paper.

II. Methodology

Before classifying, the first stage in this research is data collection using E-Nose, followed by feature extraction, normalization, and then performing feature selection with several FSAs. After that, the data can be classified by implementing a number of supervised learning methods, with an evaluation using 10-fold cross-validation as a final step. Fig. 1 depicts the research's flowchart, which is described in more detail in the following sections.

A. Data Collection

The data used in this research were taken using the E-Nose system. There are 10 output sensors used in this research, 8 of which are TGS sensors, and the other 2 are temperature and humidity sensors. The TGS sensors used in this research consisted of TGS 813, TGS 2603, TGS 2600, TGS 826, TGS 2620, TGS 832, and TGS 822. TGS 2612. The sensors used to measure temperature and humidity were DHT 21. The detailed procedure of data collection is presented in [23].

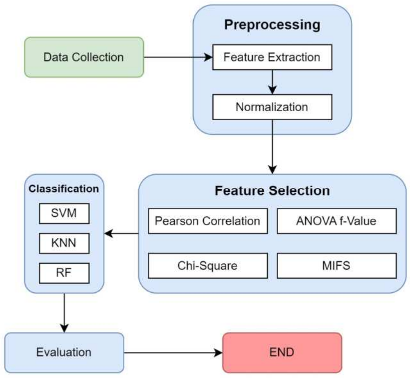

Fig. 1. Flowchart of the Research Methodology

The objects to be used in this research are 4 classes of types of meat, i.e. chicken, beef, lamb, and pork. The number of each type of meat sequentially is 108, 61, 27, and 80, so the total of all data used in this research is 276 meat sample data. The odor of the meat will be smelled by the E-Nose sensor for 6 minutes. Fig. 2 shows an example of the signals that are the output of the TGS 832 sensor on the E-Nose system. The red curve represents the lamb sample, the cyan curve represents the beef sample, the green curve represents the pork sample, and the blue curve represents the chicken sample.

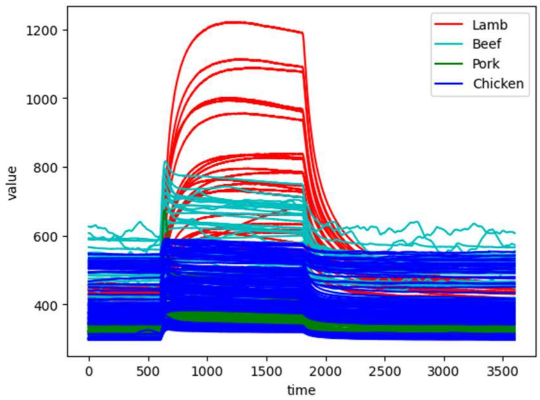

Fig. 2. Signal response from TGS 832 sensor

B. Data Preprocessing

The preprocessing stages in this research are feature extraction and normalization. The process of extracting features is implemented by utilizing two statistical parameters, i.e. average and standard deviation. Eq. (1) and Eq. (2) show the mathematical formula of average and standard deviation.

$$\bar{x} = \frac{\sum x}{n} \tag{1}$$

$$SD = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} \tag{2}$$

where x is the value of each observational data, $\bar{x}$ is the mean value of x, n is the amount of data, and SD is the standard deviation.

Because there are 10 output sensors and we use 2 statistical parameters, the total features used in this research are 20, i.e. TGS 822 mean, TGS 822 std, TGS 2612 mean, TGS 2612 std, TGS 2620 mean, TGS 2620 std, TGS 832 mean, TGS 832 std, TGS 826 mean, TGS 826 std, TGS 2603 mean, TGS 2603 std, TGS 2600 mean, TGS 2600 std, TGS 813 mean, TGS 813 std, mean temperature, temperature std, mean humidity, and humidity std.

After knowing the results of feature extraction, the next step is to normalize or standardize the data. There are many other normalization techniques, however Min-Max Scaling is one of the most widely utilized techniques in this study. As a data normalization technique, min-max scaling divides each feature's value into a range, typically ranging from 0 to 1 [24]. Eq. (3) shows the mathematical formula of Min-Max Scaling.

$$x_{new} = \frac{x-x_{min}}{x_{max}-x_{min}} \tag{3}$$

where x is the original value, x_{min} is the minimum value of x, x_{max} is the maximum value of x, and x_{new} is the result value of Min-max Scaling.

C. Feature Selection Algorithm

1) Pearson Correlation

Pearson correlation is a statistical metric that is usually used to measure the degree of linear relationship between two continuous variables [23]. The result of Pearson Correlation is expressed in the range between -1 and 1, where: if the correlation is 1, it indicates a perfectly positive linear relationship. If the correlation is -1, it shows a perfectly negative linear relationship. It is evident that there is no linear relationship between the two variables if the correlation is zero. The mathematical formula of Pearson Correlation is shown in Eq. (4).

$$r = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2}} \tag{4}$$

where X and Y with n observations. Where r is the coefficient Pearson Correlation and X_i and Y_i are value each variable i observation. Meanwhile, $\bar{X}$ and $\bar{Y}$ are the mean values of X and Y.

2) ANOVA f-Value

ANOVA is a technique used to test whether there is a significant difference between the mean values of the groups. The F-value in ANOVA is calculated by comparing the variability between groups with the variability within groups [25]. If the between-group variability is greater than within-group variability, then the F-value will be large, indicating a significant difference between at least one pair of groups. Eq. (5) shows the mathematical formula of ANOVA f-Value.

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{X_i}-\bar{X})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X_i})^2} \tag{5}$$

where F is the variability value, k is the number of groups, n_i is the number of observations in the group i, X_{ij} is the value in group i on observation j, $\bar{X_i}$ is the mean value of group i and $\bar{X}$ is the mean value of all data. In the context of feature selection, F-value ANOVA can be used in feature selection for datasets with categorical target variables, where we want to identify features that have significant differences between groups.

3) Chi-Square

Chi-square (χ^2) is a metric used to measure the relationship between two categorical variables in the form of a contingency table [26]. The contingency table is a table that shows the frequency distribution of a pair of interrelated variables. The Chi-Square value can be calculated using Eq. (6).

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (6)$$

where X^2 is the Chi-Square value, O is the observed value, and E is the expected value. Features with high Chi-Square values indicate that there is a significant relationship between these features and the target variable. Features that have high chi-square values can be selected as features to be used in machine learning models.

4) *Mutual Information Feature Selection*

Mutual Information Feature Selection (MIFS) is a method used in feature selection to measure how far information from one feature can provide insight into the target variable or other variables [27]. MIFS is a popular approach for feature selection for high-dimensional data. A feature is important and will be utilized in the model if it offers a lot of insightful data about the target variable. The MI value can be calculated using Eq. (7).

$$MI = H(X) + H(Y) - H(X, Y) \quad (7)$$

where MI is the MI value, $H(X)$ is the entropy of variable X, $H(Y)$ is the entropy of variable Y, and $H(X, Y)$ is the combined entropy of X and Y. One of the strengths of MI is MI can capture non-linear relationships between features and the target variable.

D. Classification Methods

1) *Support Vector Machine*

SVM is one of the machine learning algorithms used for classification and regression tasks. This algorithm is known for its concept of being able to perform classification and regression by finding the best hyperplane (line, plane, or hyperplane) to describe the boundary between data classes [28]. Normally, SVM will segregate classes using a line. But sometimes there is a case where the data cannot be separated linearly. To overcome this, a higher dimensional space is used that allows class separation to be performed non-linearly. This trick is commonly known as the kernel trick.

2) *K-Nearest Neighbor*

As the name implies, conceptually KNN is a method that works by finding the "nearest neighbor" of a data point to make a decision [29]. The K value in K-NN indicates the number of nearest neighbors that will be used to make decisions when classifying or predicting data. The K value can affect the results of accuracy in predicting data. In general, smaller K tends to result in more complex models and can capture small variations in the data. However, this can also make the model more susceptible to interference or noise. On the other hand, a bigger K may lose crucial data features but tends to create a more robust model and greater generalization.

3) *Random Forest*

Random forest is a supervised learning algorithm that is commonly used for data classification processes. Conceptually, the random forest is a method that uses multiple decision trees [30]. Previous research stated that by using multiple trees, the accuracy of prediction results will increase compared to using only one decision tree. First of all, this method will create many variations of subsamples, and each sub-sample will train a decision tree. After all the trees have been trained, the random forest method will decide which class the data will go into based on the majority voting of the results of each tree.

III. Result and Discussion

A. Result of Features Selection

This research used 4 FSA methods. Three combinations of parameters were applied to each method, so there were 12 FSA combinations, which can be seen in Table I.

In previous studies, the Pearson correlation has an important parameter which is commonly known as the threshold. We use thresholds of 80%, 85%, and 90% in this research. Pearson correlation produces 12 features using thresholds of 80% and 85%, and and produces 17 features using a threshold of 90%. For ANOVA F-value, Chi-Square, and MIFS, we use the same parameter values, i.e. k = 5, k = 10, and k = 15. K here means the number of features that you want to maintain. However, even though the number of features produced is the same, the features selected are still different.

TABLE I. Selected Features

FSA	***Selected Features***
Pearson Correlation (80%)	12 Features; TGS 822 mean, TGS 822 std, TGS 2612 mean, TGS 2612 std, TGS 2620 mean, TGS 2603 mean, TGS 2603 std, TGS 813 std, Temp mean, Temp std, Humid mean, Humid std.
Pearson Correlation (85%)	12 Features; TGS 822 mean, TGS 822 std, TGS 2612 mean, TGS 2612 std, TGS 2620 mean, TGS 2603 mean, TGS 2603 std, TGS 813 std, Temp mean, Temp std, Humid mean, Humid std.
Pearson Correlation (90%)	17 Features; TGS 2620 mean, TGS 832 mean, TGS 832 std, TGS 822 mean, TGS 822 std, TGS 2612 mean, TGS 2612 std, TGS 826 mean, TGS 2603 mean, TGS 2603 std, TGS 2600 std, TGS 813 mean, TGS 813 std, Temp mean, Temp std, Humid mean, Humid std.
ANOVA f-Value (k=5)	5 Features; TGS 2620 mean, TGS 832 mean, TGS 832 std, TGS 826 std, TGS 813 mean.
ANOVA f-Value (k=10)	10 Features; TGS 826 std, TGS 2603 mean, TGS 2600 mean, TGS 813 mean, TGS 813 std, TGS 2620 mean, TGS 2620 std, TGS 832 mean, TGS 832 std, TGS 826 mean.
ANOVA f-Value (k=15)	15 Features; TGS 826 mean, TGS 826 std, TGS 2603 mean, TGS 2600 mean, TGS 822 mean, TGS 822 std, TGS 2612 mean, TGS 2620 mean, TGS 2620 std, TGS 832 mean, TGS 832 std, TGS 813 mean, TGS 813 std, Temp mean, Temp std.
Chi-square (k=5)	5 Features; TGS 813 mean, TGS 813 std, TGS 2620 mean, TGS 832 std, TGS 826 std.
Chi-square (k=10)	10 Features; TGS 2603 mean, TGS 2600 mean, TGS 813 mean, TGS 813 std, TGS 2620 mean, TGS 832 mean, TGS 832 std, TGS 826 mean, TGS 826 std, Temp mean.
Chi-square (k=15)	15 Features; TGS 832 std, TGS 826 mean, TGS 826 std, TGS 822 mean, TGS 822 std, TGS 2612 mean, TGS 2620 mean, TGS 2620 std, TGS 2600 mean, TGS 813 mean, TGS 813 std, TGS 832 mean, TGS 2603 mean, Temp mean, Temp std.
MIFS (k=5)	5 Features; TGS 2620 mean, TGS 832 mean, TGS 826 mean, TGS 813 mean, Temp mean.
MIFS (k=10)	10 Features; TGS 813 mean, TGS 822 mean, TGS 832 mean, TGS 826 mean, TGS 2603 mean, TGS 2600 mean, TGS 2612 mean, TGS 2620 mean, TGS 813 std, Temp mean.
MIFS (k=15)	15 Features; TGS 2620 mean, TGS 2620 std, TGS 2603 mean, TGS 822 mean, TGS 2612 mean, TGS 832 mean, TGS 832 std, TGS 826 mean, TGS 826 std, TGS 2600 mean, TGS 813 mean, TGS 813 std, Temp mean, Temp std, Humid mean.

B. Classification Result

After selecting features using each FSA, the next step is to classify the data into four classes. In previous researches, SVM, K-NN, and RF algorithms produce predictive results that have good accuracy [31]–[33]. So, we use these three algorithms in this research. Python 3.10.12 is used for all signal processing, including classification, along with a number of related libraries including scikit-learn and matplotlib.

The first classification algorithm implemented in this research is SVM. Parameters in the SVM algorithm can have a large impact on the performance and speed of model convergence. The selection of the right parameters is very important to get optimal results [34]. Some of the main parameters that can be controlled in SVM and applied in this research are kernel, C, and gamma (γ). The values for the kernel are 'linear' and 'rbf'. Whereas for C and gamma (γ) are 0.001, 0.01, 0.1, 1, 5, 10, 100. Just like SVM, the parameters in the KNN and RF algorithms also have a big impact on model performance. The parameters for KNN are the number of neighbors and the weighting method. The values for the number of neighbors are 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10, while the values for the weighting method are 'distance' and 'uniform'. For the RF algorithm, 3 parameters are controlled, i.e. the number of estimators, maximum depth, and min_samples_split which controls the minimum number of samples needed to divide a node in a decision tree. The number of estimators parameter has the following values: 5, 10, 20, 30, 50, 100, and 200. The values for the maximum depth parameter are none, 10, 20, and 30. While the values for the min_samples_split parameter are 2, 5, 10, 20, 30, 50.

K-fold cross-validation is used to split the dataset into K subsets prior to classification. There will be two subsets used: one for training and the other for testing. In this study, K = 10 is the value that was employed. The evaluation results for various FSA combinations are displayed in Table II.

The MIFS (k=10), as can be seen in the Table II, has the highest accuracy value. To make it easier to analyze and compare, Fig. 3 displays an evaluation chart of experimental results using MIFS (k=10) combined with 3 machine learning methods. SVM generated the results with the best accuracy, which was 98.91%. These hyperparameters are used to generate the result: kernel parameters = "rbf," C = 100, and gamma (γ) = 5. The results of the confusion matrix from SVM-MIFS (k = 10) can be seen in Fig. 4.

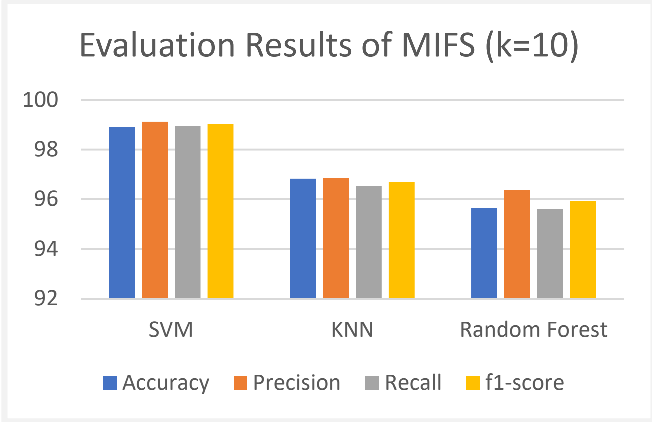

Fig. 3. Evaluation Results of MIFS (k = 10)

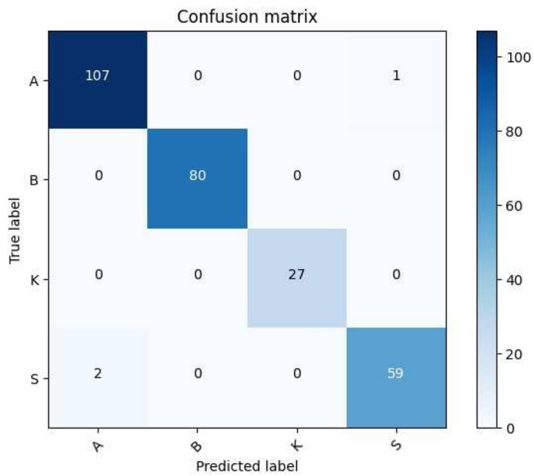

Fig. 4. Confusion Matrix of SVM-MIFS (k = 10)

TABLE II. PRECISION, RECALL, AND F1-SCORE OF FSA COMBINATION

FSA	*SVM*				*KNN*				*RF*			
	Acc	*P*	*R*	$f1-score$	*Acc*	*P*	*R*	$f1-score$	*Acc*	*P*	*R*	$f1-score$
Pearson Correlation (80%)	96.38	96.60	96.81	96.70	94.57	95.15	94.68	94.88	95.29	96.01	95.30	95.58
Pearson Correlation (85%)	96.38	96.60	96.81	96.70	94.57	95.15	94.68	94.88	95.29	96.01	95.30	95.58
Pearson Correlation (90%)	96.74	96.97	97.12	97.04	95.65	96.32	95.71	95.97	94.57	95.19	95.10	95.11
ANOVA f-Value (k=5)	90.94	92.20	91.15	91.63	88.04	90.95	89.18	89.68	88.41	90.49	89.41	89.80
ANOVA f-Value (k=10)	97.83	98.08	98.00	98.04	97.10	97.44	97.35	97.40	94.93	95.95	95.15	95.44
ANOVA f-Value (k=15)	97.46	98.01	97.49	97.73	96.01	96.76	96.03	96.32	94.93	95.84	94.98	95.31
Chi-square (k=5)	89.49	90.61	89.64	90.10	86.96	89.69	87.26	87.76	91.67	93.38	92.13	92.59
Chi-square (k=10)	97.46	98.01	97.49	97.73	97.46	97.97	97.31	97.59	96.01	96.93	96.03	96.37
Chi-square (k=15)	97.46	98.01	97.49	97.73	96.01	96.76	96.03	96.32	94.93	95.84	94.98	95.31
MIFS (k=5)	93.84	94.50	93.72	94.04	94.57	95.40	95.30	95.33	92.39	93.33	93.27	93.27
MIFS (k=10)	98.91	99.12	98.95	99.03	96.38	96.86	96.53	96.69	95.65	96.38	95.62	95.93
MIFS (k=15)	97.46	97.61	96.99	97.19	96.01	96.70	95.95	96.27	96.38	97.16	96.26	96.60

*all results in percentage(%)

C. Comparison with Previous Researches

This research also compares the results of the FSA with several similar papers that have been published earlier. There are several combinations of the FSA method and the classification method, such as ANOVA-SVM, Pearson Correlation-SVM, etc. A more detailed comparison can be seen in Table III. In the table, it can be seen that the best performance was achieved by the method used in this research with an accuracy of 98.91%.

TABLE III. COMPARISON WITH PREVIOUS RESEARCH

Author	***FSA***	***Classification Method***	***Acc (%)***
Mhd Arief Hasan, et al. [19]	Pearson Correlation and Chi-Square	KNN	82.00
Shoffi Izza Sabilla, et al. [23]	Pearson Correlation	SVM	92.00
Irzal Ahmad Sabilla, et al. [35]	ANOVA	SVM	94.12
Hannaneh Mahdavi, et al. [20]	Customize Method	SVM	98.80
This Research	MIFS	SVM	98.91

IV. CONCLUSION

In order to cope with redundancy and data overlapping in electronic nose signal data, this research provided a number of FSA techniques. Chicken, beef, lamb, and pork were correctly identified as the four types of meat from the E-Nose data. This research demonstrates that the MIFS approach with k=10 is the most effective FSA based on the performance outcomes of various combinations. TGS 813 mean, TGS 822 mean, TGS 832 mean, TGS 826 mean, TGS 2603 mean, TGS 2600 mean, TGS 2612 mean, TGS 2620 mean, TGS 813 std, and Temp mean are the characteristics that are still kept and used for the machine learning model using the MIFS approach (k=10). By applying a combination of MIFS (k = 10) with SVM, an accuracy of 98.91% is attained. Meanwhile, the value of precision is 99.12%, recall is 98.95%, and F1-score is 99.03%.

ACKNOWLEDGMENT

The funding for this research was provided by the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia as part of the PMDSU Scholarship Program managed by Institut Teknologi Sepuluh Nopember No. 3596/E4/DT.04.02/2023, and the project plan of the Publication Incentive Program and Intellectual Property Rights (PPHKI), supervised by the Institut Teknologi Sepuluh Nopember.

REFERENCES

[1] K. A. Taher, B. Mohammed Yasin Jisan, and Md. M. Rahman, "Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection," in *2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)*, IEEE, Jan. 2019, pp. 643–646. doi: 10.1109/ICREST.2019.8644161.

[2] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 56–70, May 2020, doi: 10.38094/jastt1224.

[3] Y. Wang and T. Li, "Local feature selection based on artificial immune system for classification," *Applied Soft Computing Journal*, vol. 87, Feb. 2020, doi: 10.1016/j.asoc.2019.105989.

[4] T. H. Chowdhury, K. N. Poudel, and Y. Hu, "Time-Frequency Analysis, Denoising, Compression, Segmentation, and Classification of PCG Signals," *IEEE Access*, vol. 8, pp. 160882–160890, 2020, doi: 10.1109/ACCESS.2020.3020806.

[5] W. Cao, R. Wang, M. Fan, X. Fu, H. Wang, and Y. Wang, "A new froth image classification method based on the MRMR-SSGMM hybrid model for recognition of reagent dosage condition in the coal flotation process," *Applied Intelligence*, vol. 52, no. 1, pp. 732–752, Jan. 2022, doi: 10.1007/s10489-021-02328-z.

[6] N. Cheng, A. J. Schulte, F. Santosa, and J. H. Kim, "Machine learning application identifies novel gene signatures from transcriptomic data of spontaneous canine hemangiosarcoma," *Brief Bioinform*, vol. 22, no. 4, Jul. 2021, doi: 10.1093/bib/bbaa252.

[7] S. M. Rezaeijo, M. Ghorvei, and M. Alaei, "A machine learning method based on lesion segmentation for quantitative analysis of CT radiomics to detect COVID-19," in *6th Iranian Conference on Signal Processing and Intelligent Systems, ICSPIS 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/ICSPIS51611.2020.9349605.

[8] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, "A Chi-Square Statistics Based Feature Selection Method in Text Classification," in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, IEEE, Nov. 2018, pp. 160–163. doi: 10.1109/ICSESS.2018.8663882.

[9] I. M. Nasir *et al.*, "Pearson correlation-based feature selection for document classification using balanced training," *Sensors (Switzerland)*, vol. 20, no. 23, pp. 1–18, Dec. 2020, doi: 10.3390/s20236793.

[10] P. Huang, X. Tan, C. Chen, X. Lv, and Y. Li, "AF-SENet: Classification of cancer in cervical tissue pathological images based on fusing deep convolution features," *Sensors (Switzerland)*, vol. 21, no. 1, pp. 1–20, Jan. 2021, doi: 10.3390/s21010122.

[11] R. Sarno, K. Triyana, S. I. Sabilla, D. R. Wijaya, D. Sunaryono, and C. Fatichah, "Detecting Pork Adulteration in Beef for Halal Authentication using an Optimized Electronic Nose System," *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.3043394.

[12] W. Jia, G. Liang, Z. Jiang, and J. Wang, "Advances in Electronic Nose Development for Application to Agricultural Products," *Food Analytical Methods*, vol. 12, no. 10. Springer New York LLC, pp. 2226–2240, Oct. 01, 2019. doi: 10.1007/s12161-019-01552-1.

[13] M. Nomura, E. Osada, T. Tokita, T. Iwamoto, and Y. Manome, "Measurement and differentiation of banana juice scent using an electronic nose FF-2A," *PeerJ*, vol. 9, Jan. 2021, doi: 10.7717/peerj.10638.

[14] S. I. Sabilla and R. Sarno, "Development of wavelet transforms to predict methane in chili using the electronic nose," in *2017 International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA)*, IEEE, Oct.

2017, pp. 271–276. doi: 10.1109/ICAMIMIA.2017.8387600.

[15] Hariyanto, R. Sarno, and D. R. Wijaya, "Detection of diabetes from gas analysis of human breath using e-Nose," in *2017 11th International Conference on Information & Communication Technology and System (ICTS)*, IEEE, Oct. 2017, pp. 241–246. doi: 10.1109/ICTS.2017.8265677.

[16] Y. A. Sihombing and M. Alfarandi Ritonga, "Detection of the cooking oil aroma by using a gas sensor in an electronic nose system," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, 2021. doi: 10.1088/1742-6596/1811/1/012105.

[17] R. Sarno and D. R. Wijaya, "Recent development in electronic nose data processing for beef quality assessment," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 17, no. 1, pp. 337–348, Feb. 2019, doi: 10.12928/TELKOMNIKA.v17i1.10565.

[18] D. R. Wijaya and F. Afianti, "Stability Assessment of Feature Selection Algorithms on Homogeneous Datasets: A Study for Sensor Array Optimization Problem," *IEEE Access*, vol. 8, pp. 33944–33953, 2020, doi: 10.1109/ACCESS.2020.2974982.

[19] M. A. Hasan, R. Sarno, and S. I. Sabilla, "Optimizing machine learning parameters for classifying the sweetness of pineapple aroma using electronic nose," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 5, pp. 122–132, Oct. 2020, doi: 10.22266/ijies2020.1031.12.

[20] H. Mahdavi, S. Rahbarpour, S. M. Hosseini-Golgoo, and H. Jamaati, "Exploring Informative Response Features of Two Temperature Modulated Gas Sensors at a Wide Range of Relative Humidity," in *2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, IEEE, Dec. 2021, pp. 01–05. doi: 10.1109/ICSPIS54653.2021.9729343.

[21] M. Xu, J. Wang, and L. Zhu, "Tea quality evaluation by applying E-nose combined with chemometrics methods," *J Food Sci Technol*, vol. 58, no. 4, pp. 1549–1561, Apr. 2021, doi: 10.1007/s13197-020-04667-0.

[22] L. Y. Chen, C. C. Wu, T. I. Chou, S. W. Chiu, and K. T. Tang, "Development of a dual MOS electronic nose/camera system for improving fruit ripeness classification," *Sensors (Switzerland)*, vol. 18, no. 10, Oct. 2018, doi: 10.3390/s18103256.

[23] S. I. Sabilla, R. Sarno, and K. Triyana, "Optimizing threshold using pearson correlation for selecting features of electronic nose signals," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 6, pp. 81–90, 2019, doi: 10.22266/ijies2019.1231.08.

[24] S. Gupta and A. Chug, "Software maintainability prediction of open source datasets using least squares support vector machines," *Journal of Statistics and Management Systems*, vol. 23, no. 6, pp. 1011–1021, Aug. 2020, doi: 10.1080/09720510.2020.1799501.

[25] P. Mishra, U. Singh, C. Pandey, P. Mishra, and G. Pandey, "Application of student's t-test, analysis of variance, and covariance," *Ann Card Anaesth*, vol. 22, no. 4, p. 407, 2019, doi: 10.4103/aca.aca_94_19.

[26] F. Mahan, M. Mohammadzad, S. M. Rozekhani, and W. Pedrycz, "Chi-MFlexDT:Chi-square-based multi flexible fuzzy decision tree for data stream classification," *Appl Soft Comput*, vol. 105, Jul. 2021, doi: 10.1016/j.asoc.2021.107301.

[27] W. Ke, C. Wu, Y. Wu, and N. N. Xiong, "A new filter feature selection based on criteria fusion for gene microarray data," *IEEE Access*, vol. 6, pp. 61065–61076, 2018, doi: 10.1109/ACCESS.2018.2873634.

[28] M. A. Chandra and S. S. Bedi, "Survey on SVM and their application in image classification," *International Journal of Information Technology (Singapore)*, vol. 13, no. 5, Oct. 2021, doi: 10.1007/s41870-017-0080-1.

[29] H. Henderi, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," *IJIIS: International Journal of Informatics and Information Systems*, vol. 4, no. 1, pp. 13–20, Mar. 2021, doi: 10.47738/ijiis.v4i1.73.

[30] T. C. Amri, R. Sarno, R. Abdillah, F. A. Haq, A. F. Septiyanto, and D. Sunaryono, "Filter Validation for Detecting Outliers of Photoplethysmograph Data," in *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, IEEE, Feb. 2023, pp. 012–017. doi: 10.1109/ICAIIC57133.2023.10067110.

[31] N. Dogru and A. Subasi, "Traffic accident detection using random forest classifier," in *2018 15th Learning and Technology Conference (L&T)*, IEEE, Feb. 2018, pp. 40–45. doi: 10.1109/LT.2018.8368509.

[32] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies (Basel)*, vol. 9, no. 3, Sep. 2021, doi: 10.3390/technologies9030052.

[33] F. A. I. Achyunda Putra, F. Utaminingrum, and W. F. Mahmudy, "HOG Feature Extraction and KNN Classification for Detecting Vehicle in The Highway," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 14, no. 3, p. 231, Jul. 2020, doi: 10.22146/ijccs.54050.

[34] I. Aljarah, A. M. Al-Zoubi, H. Faris, M. A. Hassonah, S. Mirjalili, and H. Saadeh, "Simultaneous Feature Selection and Support Vector Machine Optimization Using the Grasshopper Optimization Algorithm," *Cognit Comput*, vol. 10, no. 3, pp. 478–495, Jun. 2018, doi: 10.1007/s12559-017-9542-9.

[35] I. A. Sabilla, Z. A. Cahyaningtyas, R. Sarno, A. Al Fauzi, D. R. Wijaya, and R. Gunawan, "Classification of Human Gender from Sweat Odor using Electronic Nose with Machine Learning Methods," in *Proceedings - 2021 IEEE Asia Pacific Conference on Wireless and Mobile, APWiMob 2021*, Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 109–115. doi: 10.1109/APWiMob51111.2021.9435205.