Comparison of Supervised Learning Algorithms for Cigarette and Vape Smoke Classification using Electronic Nose

Muhammad Agung Farghani, Nurul Izzah Wijayakusuma, and Budi Sumanto*
Department of Electrical Engineering and Informatics, Vocational College, Universitas Gadjah Mada, Jl. Yacaranda Sekip Unit III, Yogyakarta, Indonesia 55281

Abstract: This research discusses applying the Supervised Learning method using Electronic Nose to classify the types of cigarette and vape smoke in the air. Electronic Nose is used as a scent detector that can identify the characteristics of smoke from both sources. Three Supervised Learning algorithms, namely KNN, SVM, and Decision Tree, were applied to compare the performance in classifying smoke types. The data comprised reference air samples, air contaminated by manufactured cigarette smoke, rolled cigarettes, and vape. The results showed that all three Supervised Learning algorithms successfully provided an excellent classification for cigarette and vape smoke types using data from Electronic Nose. The best accuracy result was achieved by SVM, with an accuracy rate of 96.55%. This research contributes to identifying sources of air pollution that have the potential to endanger human health.

Keywords: Air; electronic Nose; smoke; supervised learning

*Corresponding author: budi.sumanto@ugm.ac.id

http://dx.doi.org/10.12962/j24604682.v20i3.17939
2460-4682

I. INTRODUCTION

Air is one of the vital elements that are beneficial for the survival of human life and ecosystems. Usually, gas in the air is composed of 78.1% nitrogen, 20.93% oxygen, 0.03% carbon dioxide, and other gases. If there is a change from the average composition or the air is contaminated by gas particles that endanger human health and the environment, it can be said that the air is polluted. The environment, the air is polluted. [1].

Air pollution can occur naturally and be caused by human activities, but quantitatively higher pollution is due to human activities such as transportation, industry, waste decomposition, combustion, and households. Cigarette combustion, both conventional and e-cigarettes (vapes), is one of the main contributors to air pollution that contributes to the pollution problem. [2].

Conventional cigarette combustion smoke contains various free radicals, including carbon monoxide, carbon dioxide, oxides of nitrogen, and hydrocarbon compounds. In addition, conventional cigarette smoke contains particulate components, including tar, nicotine, benzopyrene, phenol, and cadmium [3]. On the other hand, e-cigarettes (vapes) also produce vapor that is considered an air pollutant, as it contains formaldehyde, carbon monoxide, and carcinogenic compounds that can harm the body's health [4]. The conventional method used in detecting hazardous air is smelling the aroma, but this method has risks that can harm the human body [5]. In addition, humans have a limited sense of smell and need help to detect harmful gas compounds accurately. Sensor technology has created an electronic nose that can handle these problems. An electronic nose or E-nose is a system designed to detect and identify chemical compounds through odor with the help of chemical sensors so that a unique response pattern is obtained [6]. The response generated on the E-nose can be electrical voltage patterns, resistance, and unit values representing each aroma [7], this response pattern can be analyzed and used to classify the type of smoke produced from burning conventional and e-cigarettes (vape), which can be one of the efforts to control air pollution and protect public health. Machine learning algorithms have become effective in various science and technology applications, including data analysis and pattern recognition. Machine learning algorithms SVM, KNN, and Decision Trees can be used in data analysis and pattern recognition aimed at data classification. [6], [8-12].

Based on the problems and descriptions described, this study aims to classify the type of smoke produced from burning conventional and electric cigarettes (vape) using an E-nose. Classification of cigarette and vape smoke types is essential because both impact the environment and human health, thus requiring different approaches to handling human exposure. This classification can generate benefits such as air pollution control, regulatory enforcement, and the development of detection and monitoring technologies. From previous research, several algorithms such as Decision Tree, SVM, and KNN are often applied in various cases and have good performance, such as in the classification of air pollution index in DKI Jakarta; Decision Tree has an accuracy rate of 99.80% [8], on the classification of dangerous and flammable gases, KNN has an accuracy rate of 99.76% [13], and in de-

(a) (b) (c)

FIG. 1: (a) Vape, (b) Rolled cigarettes, (c) Factory cigarettes.

FIG. 2: E-nose device.

termining majors in vocational schools, the SVM algorithm has an accuracy rate of 89% [9]. Therefore, this study aims to compare the three algorithms and find the algorithm that has the best accuracy in classifying the type of smoke produced from burning cigarettes and vapes.

II. METHOD

A. Research Materials

The materials used as research objects include manufactured cigarettes, roll-up cigarettes, and vape shown in Fig. 1. Each type contains different ingredients and concentrations of substances. Conventional cigarettes, such as manufactured cigarettes and roll-up cigarettes, are primarily made from Nicotiana plants, which contain nicotine and tar, while vapes use liquid that also contains nicotine. The combustion of cigarettes produces pollutants such as carbon monoxide, carbon dioxide, and over 200 other chemical toxins [14]. Similarly, vape vapor contains hazardous compounds like formaldehyde, carbon monoxide, and carcinogenic substances that can negatively affect health [4]. In the research, the smoke produced from each object will be taken and used as samples of air conditions. The air conditions used as samples include reference air, air when given vape smoke, air when given rolling cigarette smoke, and air when given factory cigarette smoke.

TABLE I: Types of MOS Sensors and Objects Detected.

Sensor Type	Object Detection
TGS-2611	Methane, Ethanol, Isobutane, Hydrogen
TGS-2600	Air quality (Methane, Carbon Monoxide, Isobutane, Ethanol, Hydrogen, etc.)
TGS-822	Solvent vapor (Ethanol, Benzene, Acetone, Methane, Isobutane, Carbon monoxide, etc.)
TGS-813	Methane, Propane, Butane, Carbon Monoxide and Hydrogen
TGS-2602	Ammonia, 2S, toluene
MQ-135	Air Quality (Carbon Dioxide, Carbon monoxide, Ammonia, Alcohol, Toluene, Acetone, etc.)
MQ-9	Carbon Monoxide, LPG (Propane and Butane) and Methane
MQ-3	Alcohol and Benzene
MQ-137	Amonia

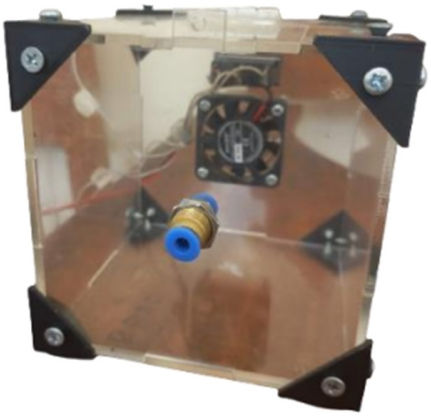


FIG. 3: Room Prototype.

B. Research Tools

In addition to the materials used, there are supporting equipment in the research process, including e-nose devices, room prototypes, and computers. The e-nose device used is shown in Fig. 2. The e-nose device consists of several main components, namely a 12V solenoid valve, 12V air pump, 4-channel relay, buck converter module, Arduino Mega Pro, and a sensor room containing nine gas sensors. The nine MOS sensors used include TGS-2611, TGS-2600, TGS-822, TGS-813, TGS-2602, MQ-135, MQ-9, MQ-3, and MQ-137, which in Table I shows the types of Metal Oxide Semiconductor sensors and the objects they detect. All actuators, such as the 12V solenoid valve and 12V air pump, get their power source directly from the 12V 2A adapter, with their conditions controlled by Arduino through relays. Meanwhile, all gas sensors and relays get a 5V voltage supply from the buck converter connected to the 12V 2A adapter. The e-nose device is used for data acquisition of reference air, air with cigarette smoke,

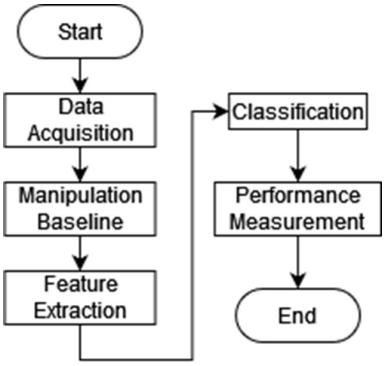


FIG. 4: Research Stages.

and air with vape smoke. The form of the room prototype is shown in Fig. 3. The prototype room samples air conditions during the data acquisition or dataset collection process.

C. Research Stages

The research is carried out in several stages, as shown in the flowchart in Fig. 4. According to the flowchart, the research follows six stages: data acquisition, baseline manipulation, feature extraction, classification, and model performance measurement. During the data acquisition stage, the e-nose device is connected to a computer equipped with a Graphical User Interface (GUI), which is shown in Fig. 5. This GUI functions to visualize data in real-time during the reading of the Metal Oxide Semiconductor (MOS) sensors. Additionally, the GUI saves the sensor reading data into a CSV file for further processing. The processes of data manipulation, feature extraction, classification, and model performance measurement are performed separately on the computer after the data has been stored in the CSV format.

D. Data Acquisition

In the data acquisition stage, the process of taking reference air data, air when given vape smoke, air when given cigarette smoke, and air when given factory cigarette smoke in the prototype room will be carried out. Measurement of one air condition will be measured 36 times, which will be tested for 100 seconds in one repetition. Repetition is done to generate a large enough dataset for the machine learning model to learn patterns better. In addition, the repetition of measurements aims to ensure that the e-nose system provides reliable and valid results in detecting cigarette smoke, as well as ensuring that the results obtained are consistent and not caused by chance or random variation. The results of testing or measuring air conditions will be stored in a file in CSV format. Each air condition sample in the dataset will be named with a label described in Table II.

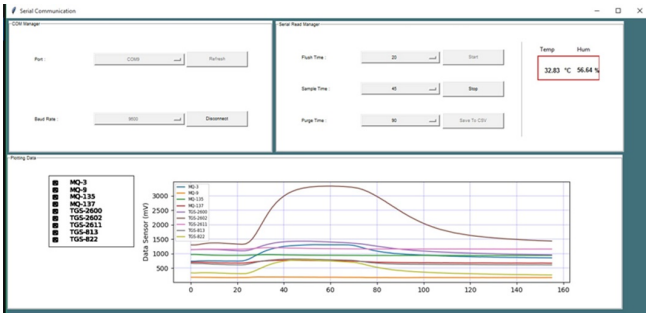


FIG. 5: Graphical User Interface.

TABLE II: Sample Dataset Label.

Sample	Label
Air Reference	UD
Air with vape smoke	UV
Air With Rolling Cigarette Smoke	URL
Air With Factory Cigarette Smoke	URP

E. Baseline Manipulation

Baseline manipulation is a preprocessing stage of sensor signals to correct inconsistent sensor values under the same conditions [15]; baseline manipulation aims to equalize the starting point of each sensor in the electronic nose system. One of the baseline manipulation methods that can be performed is the differential method, which can be used to reduce or eliminate noise caused by external factors [15]. The baseline manipulation equation is shown in Equation (1).

$$X_{ij} = V_{ij}max - V_{ij}min \tag{1}$$

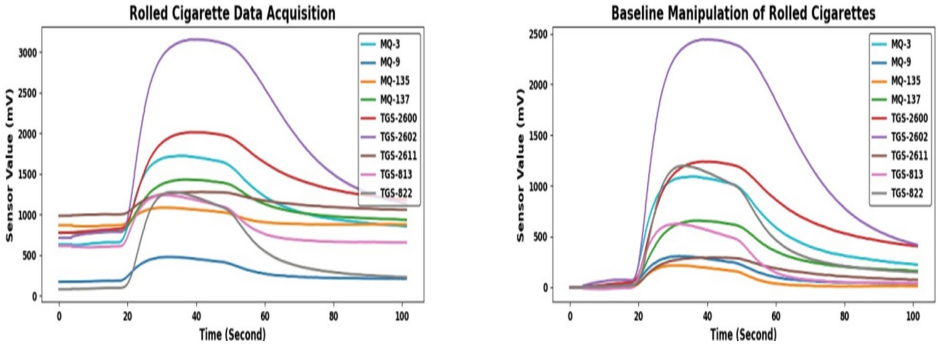
X_{ij} is the difference between two values in the dataset to be calculated, V_{ij} max is the most significant value of the entire dataset, and V_{ij} min is the smallest value. [16].

F. Feature Extraction

Feature extraction is a stage where the raw data generated by the sensor will be characterized in a more concise feature space so that optimal recognition results are obtained. In this research, feature extraction is taken from literature studies that have been applied to cigarette smoke and vape datasets by feature selection so that six features are obtained that have the slightest error value, namely mean, maximum value, variance, first frequency spectrum power, third frequency spectrum power, and fourth frequency spectrum power [17].

G. Classification

The classification process aims to classify objects based on their labels or classes. Some methods can be used in the clas-



(a) (b)

FIG. 6: (a) Data Acquisition (b) Baseline Manipulation.

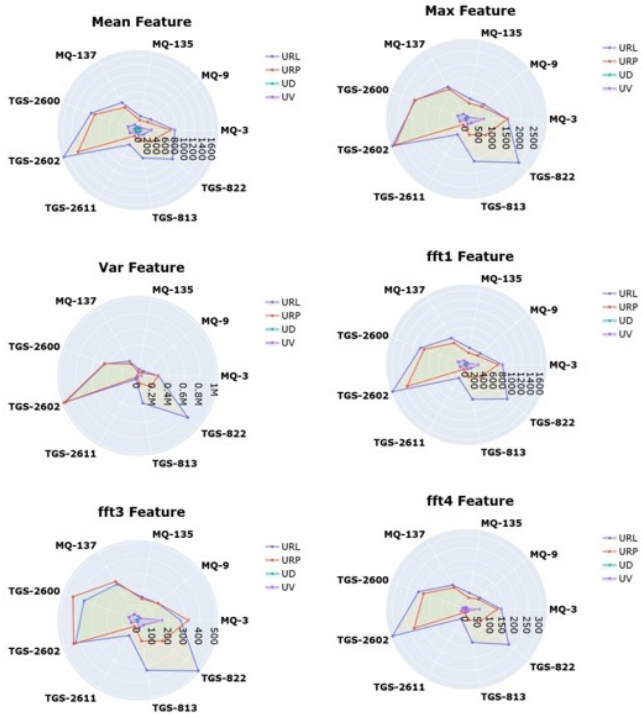


FIG. 7: Radar Plot of Feature Extraction.

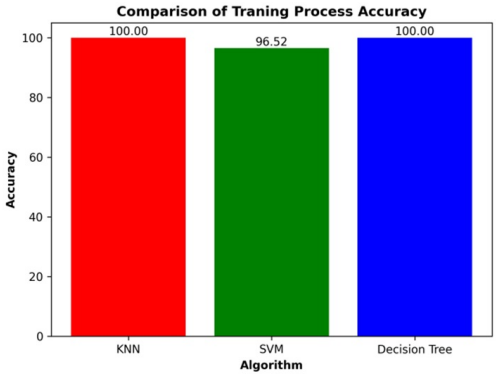


FIG. 8: Accuracy of Model Training Process.

sification process; determining the classification method depends on the characteristics of the resulting dataset. At this stage, the SVM, KNN, and Decision Tree algorithm methods will be used to classify the type of smoke produced from burning cigarettes and vapes.

H. Performance Measurement

Confusion matrices can be used to measure the performance of each classification model. The confusion matrix contains prediction results representing the obtained TP, FP, TN, and FN values and can be used to calculate the parameter values of accuracy, recall, precision, and F1-Score [18]. This parameter value can be used as an illustration of the model's performance in recognizing all existing classes. To get the accuracy, precision, and recall values using the confusion matrix, Eq. (2), Eq. (3), and Eq. (4) can be used as follows.

$$\text{Accuracy} = \frac{TP + TN}{Total} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Re call} = \frac{TP}{FN + TP} \tag{4}$$

where *TP* (True Positive) is a correct prediction and is in the positive class, *TN* (True Negative) is a correct prediction and is in the negative class, *FP* (False Positive) is a wrong prediction and is in the positive class, and *FN* (False Negative) is a wrong prediction and is in the negative class.

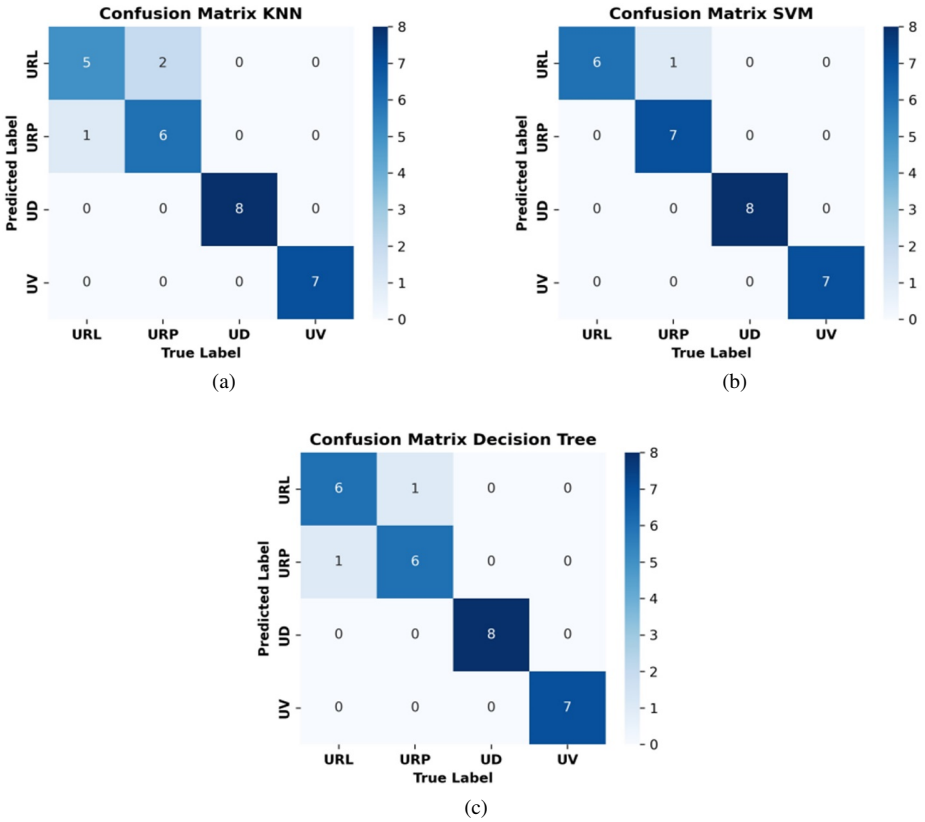


FIG. 9: (a) Confusion Matrix KNN, (b) Confusion Matrix SVM, (c) Confusion Matrix Decision Tree.

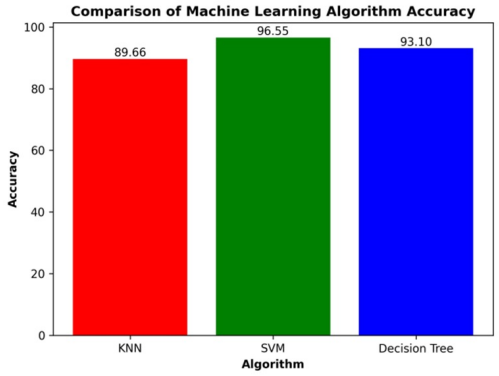


FIG. 10: Accuracy of Model Training Process.

III. RESULTS AND DISCUSSION

A. Results of Data Acquisition and Baseline Manipulation

The results obtained from air condition sampling will be stored in a file in CSV format and are still raw data that needs to be processed. An illustration of data acquisition and baseline manipulation is shown in Fig. 6. Fig. 6 shows an example of the results of data acquisition and baseline manipulation of air condition data with the first loop of roll-up cigarettes smoke illustrated in a graph plot. The x-axis is the time of the data capture process, and the y-axis is the reading value of each sensor in units of (mV). In the early stages of data acquisition, there were different variations in values between sensor readings at various times. To equalize the starting point of the data, baseline manipulation is performed by aligning so that all data starts from the same point on the y-axis at a specific initial time. The baseline value of the sensor readings, which is close to zero, indicates that the sensor readings have a low noise level.

B. Feature Extraction Results

After obtaining the same baseline value, all data is characterized in a more compact and representative feature space. From the feature extraction process, data or information on the value of six independent features for each sensor is obtained, and overall the total datasets generated is 54 features. Each feature obtained can be visualized in a radar plot so that a comparison of each class or sample's sensor feature values can be seen, shown in Fig. 7.

With radar plot visualization, the comparison of sensor

feature values from each class or sample can be seen. From the results of the radar plot shown, it can be seen that the class or sample with the most significant area is the air condition with cigarette smoke, indicating that the sample has more significant variation or higher values in certain features or attributes. Conversely, the fresh air sample has the smallest area, indicating less variation in the measured features or smaller values in certain features or attributes.

C. Classification Results, Model Performance Measurement, and Comparison

Before creating a model, the dataset will be divided into two parts: train and test data. This study has 144 datasets divided into 115 train data and 29 test data. To train the system or model, train data will be used for each algorithm. The accuracy of the SVM, KNN, and Decision Tree models during the training process is shown in Fig. 8. From the results obtained, it can be seen that the KNN and Decision Tree models are susceptible to the train data, which has an accuracy rate of 100%, while the SVM model has an accuracy rate of 96.52%. After the model has been built or has gone through the training process, it will be tested and evaluated for its performance level using the confusion matrix shown in Fig. 9. In the KNN model testing process as a whole, there are three wrong predictions; for the SVM algorithm, there is one wrong prediction, and for Decision Tree, there are two wrong predictions. With the confusion matrix, the performance or accuracy of each model can be obtained, as shown in Figure 10. The algorithm with the highest accuracy level in classifying cigarette and vape combustion smoke during testing is SVM, with an accuracy value of 96.55%. The second highest accuracy rate is the Decision Tree, with a value of 93.1%, followed by the KNN algorithm, with a value of 89.66%.

IV. CONCLUSION

This research implements Supervised Learning methods using Electronic Nose to classify cigarette and vape smoke types. Three Supervised Learning algorithms were considered: KNN, SVM, and Decision Tree. From the research, the SVM model has the highest accuracy rate of 96.55% in classifying cigarette and vape combustion smoke on the dataset used. In contrast, the KNN and Decision Tree models have lower accuracy, namely 89.66% and 93.10%. Although the KNN and Decision Tree models achieved 100% accuracy on the training data, they tend to need to be more balanced and more able to generalize to new data. Therefore, SVM is the best choice in classifying cigarette and vape combustion smoke in this study. However, it should be noted that this study has several limitations, including limitations in the number and variety of data samples used. Using a larger and more diverse dataset could improve the generalization and reliability of the results. In addition, considering the application of other algorithms or combining multiple algorithms may be a consideration for future research to improve performance and accuracy further.

Acknowledgments

To the VENOSE UGM 2023 Research Team, I would like to express my gratitude for your dedication and hard work. The great collaboration of each team member has brought this research to success. Thank you, Ilham, Ikrima, Alfonzo, and Abel, for your outstanding contributions. Every step and effort you made had a positive impact on the final outcome of this paper. Thank you for working together as a solid team, supporting each other, and facing challenges together. You have proven that good teamwork can achieve amazing results. Hopefully, this paper can benefit the development of science and make a meaningful contribution to the field. Once again, thank you for all your efforts and dedication.

[1] M. Sari, D.N. Santi, and I. Chahaya, "Analisa Kadar CO dan NO_2 di Udara dan Keluhan Gangguan Saluran Pernapasan pada Pedagang Kaki Lima di Pasar Sangkumpal Bonang Kota Padangsidimpuan," Lingkungan dan Keselamatan Kerja, vol. 3, no. 1, 2013.

[2] R.D. Ratnani, "Teknik Pengendalian Pencemaran Udara Yang Diakibatkan oleh Partikel," Majalah Ilmiah Momentum, vol. 4, no. 2, p. 2732, 2008.

[3] Afiana Rohmani, Noor Yazid, dan Aulia A. Rahmawati, Rokok Elektrik dan Rokok Konvensional Merusak Alveolus Paru, Prosiding Seminar Nasional Unimus, vol. 1, hlm. 27-32, 2018. https://prosiding.unimus.ac.id/index.php/semnas/article/view/21

[4] A. Sabir, M. Asikin, dan I. Willem, The Influence of Electric Cigarette Vapor on Ambien Air Quality in Electric Cigarette User Environment in Parepare City, Jurnal Ilmiah Manusia Dan Kesehatan, vol. 2, no. 3, hlm. 447-458, 2019. https://doi.org/10.31850/makes.v2i3.190

[5] D. Lelono and K. Prastya, "Karakterisasi Pola dan Konsentrasi Gas Polutan Berbasis E-Nose," IJEIS, vol. 3, no. 1, hlm. 8394, 2013.

[6] O. Attallah and I. Morsi, "An electronic nose for identifying multiple combustible/harmful gases and their concentration levels via artificial intelligence," Measurement (Lond), vol. 199, Aug. 2022, doi: 10.1016/j.measurement.2022.111458.

[7] K. Kusairi, M. Muthmainnah, Imam Tazi, and Moh. Fajrul Falah, "Klasifikasi Pola Aroma Teh Hijau Menggunakan Hidung Elektronik (E-Nose) Berbasis Linear Diskriminan Analisis (LDA)," JURNAL PENDIDIKAN MIPA, vol. 12, no. 3, hlm. 868874, Sep. 2022, doi: 10.37630/jpm.v12i3.682.

[8] S.S.A. Umri, M.S. Firdaus, and A. Primajaya, "Analisis Dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara Di Dki Jakarta," Jurnal Informatika dan Komputer) Akreditasi KEMENRISTEKDIKTI, vol. 4, no. 2, 2021, doi: 10.33387/jiko.

[9] N.A. Sinaga, R. Ramadani, K. Dalimunthe, M.S.A.A. Lubis, and R. Rosnelly, "Komparasi Metode Decision Tree, KNN, dan SVM Untuk Menentukan Jurusan Di SMK," Jurnal Sistem Komputer dan Informatika (JSON), vol. 3, no. 2, hlm. 94, Dec. 2021, doi: 10.30865/json.v3i2.3598.

[10] R.A. Nugraha, E.W. Hidayat, N.I. Kurniati, dan R.N. Shofa, "Klasifikasi Jenis Buah Jambu Biji Menggunakan Algoritma Principal Component Analysis dan K-Nearest Neighbor," 2023. doi: https://doi.org/10.29407/gj.v7i1.17900.

[11] A. Chandra, Gregorius, M.S.J. Immanuel, A.A.S. Gunawan, and Anderies, "Accuracy Comparison of Different Machine Learning Models in Phishing Detection," in ICOIACT 2022 - 5th International Conference on Information and Communications Technology: A New Way to Make AI Useful for Everyone in the New Normal Era, Proceeding, Institute of Electrical and Electronics Engineers Inc., 2022, p. 2429. doi: 10.1109/ICOIACT55506.2022.9972107.

[12] D.R. Prehanto, A.D. Indriyanti, I.K.D. Nuryana, and G.S. Permadi, "Classification based on K-Nearest Neighbor and Logistic Regression method of coffee using Electronic Nose," IOP Conf Ser Mater Sci Eng, vol. 1098, no. 3, p. 032080, Mar. 2021, doi: 10.1088/1757-899x/1098/3/032080.

[13] O. Attallah, and I. Morsi, "An electronic nose for identifying multiple combustible/harmful gases and their concentration levels via artificial intelligence," Measurement (Lond), vol. 199, Aug. 2022, doi: 10.1016/j.measurement.2022.111458.

[14] I. Vidiasari Aristawati, and U. Nurbaiti, "Uji kadar CO, CO 2 dan HC pada pembakaran rokok konvensional tanpa pengaruh udara luar dengan Outomotive Emission Analyzer," 2021. [Online]. Available: http://ejurnal.mipa.unsri.ac.id/index.php/jps/index

[15] D.K. Agustika, and D.K. Triyana, "The Method Of Baseline Manipulation To Overcome The Sensor Drift On Gas Sensor Test For Herbal Drinks Discrimination," J. Sains Dasar, vol. 5, no. 1, pp. 5256, 2016, doi: 10.21831/jsd.v5i1.12667.

[16] J. Yan, *et al.*, "Feature Extraction from Sensor Data for Detection of Wound Pathogen Based on Electronic Nose," Sensors and Materials, vol. 24, no. 2, p. 5773, 2012.

[17] A.A. Nugroho, W. Wijaya, J. Hendry, And B. Sumanto, "Seleksi Fitur Aroma Teh Kombucha menggunakan ANN untuk Optimasi Kinerja Sistem E-nose," ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, and Teknik Elektronika, vol. 10, no. 2, hlm. 334, Apr. 2022, doi: 10.26760/elkomika.v10i2.334.

[18] I. Widhi Saputro, and B. Wulan Sari, "Uji Performa Algoritma Nave Bayes untuk Prediksi Masa Studi Mahasiswa Nave Bayes Algorithm Performance Test for Student Study Prediction," Citec Journal, vol. 6, no. 1, 2019.