ComTech: Computer, Mathematics and Engineering Applications, 15(2), December 2024, 109−118 P-ISSN: 2087-1244
DOI: 10.21512/comtech.v15i2.12051 E-ISSN: 2476-907X

A Novel Machine Learning for Ethanol and Methanol Classification with Capacitive Soil Moisture (CSM) Sensors

Devina Intan Sari[1], Suryasatriya Trihandaru[2], and Hanna Arini Parhusip[3]*

[1-3]Magister Sains Data, Fakultas Sains & Matematika, Universitas Kristen Satya Wacana
Salatiga, Indonesia 50711
[1]632023001@student.uksw.edu, [2]suryasatriya@uksw.edu, [3]hanna.parhusip@uksw.edu

Received: 19th August 2024/ **Revised**: 5th November 2024/ **Accepted**: 6th November 2024

How to Cite: Sari, D. I., Trihandaru, S., & Parhusip, H. A. (2024). A Novel Machine Learning for Ethanol and Methanol Classification with Capacitive Soil Moisture (CSM) Sensors. *ComTech: Computer, Mathematics and Engineering Applications, 15*(2), 109−118. https://doi.org/10.21512/comtech.v15i2.12051

Abstract **-** Although Gas Chromatography (GC) is highly accurate, it is costly, highlighting the need for a more affordable method for alcohol detection. Ethanol and methanol have different evaporation rates and dielectric constants, suggesting the potential for classification as an alternative initial step to GC based on differences in dielectric due to evaporation using Capacitive Soil Moisture (CSM) sensors, although it has not been previously attempted. The research aimed to present a novel machine learning for ethanol and methanol classification with CSM sensors. The method involved placing evaporated samples on CSM plates and measuring the change in evaporative dielectric properties over time. The data were then processed using Python, preprocessing data, splitting data, and training various classifiers with key differentiators based on standard deviation, mean, difference, and cumulative summary. Then, model accuracy was evaluated. The research results show that the approach can distinguish between pure ethanol and methanol based on the dielectric differences in each substance's evaporation rate using machine learning training methods with classifiers such as Random Forest, Extra Trees, Gaussian Naive Bayes, AdaBoost, and Logistic Regression with seven folds in cross-validation, L2 regularization, and Newton-Cholesky solver, with accuracies of 96.67%, 96.67%, 96.67%, 93.33%, and 93.33%, respectively. Although the research is limited to the classification of two types of alcohol, the novel approach can classify methanol and ethanol, leading to a potential initial step in determining alcohol content in the future. It can be an alternative to GC with a simpler and more affordable setup using CSM sensors.

Keywords**:** machine learning, ethanol classification, methanol classification, Capacitive Soil Moisture (CSM)

I. INTRODUCTION

Different types of alcohol, particularly ethanol and methanol, which are widely used and circulated in society, both in the health sector and industry, such as in fuel mixtures affecting engine performance, have different physical and chemical properties (Iliev, 2021). Ethanol has a boiling point of 78.4°C, while methanol has a boiling point of 64.5°C (Yanti et al., 2019). Ethanol and methanol also have differences in viscosity due to the difference in molecular mass. Ethanol has a higher molecular mass than methanol, thus making ethanol more viscous than methanol (Putri & Kalsi, 2017). It is a key factor influencing the evaporation rate of the substances. Each type of substance has a different dielectric constant, such as 24.30 for ethanol, 22.60 for methanol, and 80.40 for water, with a higher dielectric constant indicating greater polarity of the compound (Septiana & Asnani, 2013). Each mixture of solutions, such as ethanol and methanol, has different dielectric properties due to molecular interactions within the ethanol and methanol mixture and changes in dielectric structure and properties as the ethanol content in the methanol compound increases (Lone et al., 2008).

Ethanol and methanol are commonly distinguished accurately using Gas Chromatography (GC). Additionally, GC determines the concentration

*Corresponding Author

of volatile compounds, such as in the analysis of formic acid, methanol, and the quantification of ethanol and methanol from bioreactor samples using Gas Chromatography-Flame Ionization Detection (GC-FID) (Joseph et al., 2022) and determining the chemical composition of alcoholic beverages using Gas Chromatography-Mass Spectrometry (GC-MS) (Savchuk et al., 2020) and GC-FID (Paolini et al., 2022). Ethanol can be used as an internal standard for quantifying volatile compounds in alcoholic products with GC-MS (Korban et al., 2021). However, procuring GC requires high costs and a long time from powering up the device to obtaining measurements. GC also requires a relatively expensive gas supply that must be replaced periodically when depleted. Therefore, there is a need for a more affordable alternative to determining alcohol content, as proposed in the research.

Based on the differences in the physical and chemical properties of various types of alcohol, specifically ethanol and methanol, the differences in evaporation rates related to changes in dielectric properties have the potential to be distinguishing factors between these two types of alcohol. One Arduino sensor, the Capacitive Soil Moisture (CSM) sensor, typically used to determine soil moisture, operates based on the principle of dielectric properties and has the tendency to differentiate a solution when it has different dielectric properties (Hrisko, 2020). The differences in dielectric constants for various types of alcohol have been previously studied using the Kirkwood model calculation with a capacitor under a pressure of 01.3 kPa at temperatures of 283.15 and 293.15 K. The working principle of the instrument involves measuring the change in electrical signal in a capacitor and resistor related to the dielectric constant of the liquid sample tested, allowing the determination of differences in dielectric constants in various liquids such as water, ethanol, methanol, and butanol (Mohsen-Nia et al., 2010).

The classification of compounds, in this case, ethanol or methanol, can be performed using machine learning methods, which is a branch of artificial intelligence that enables the development of algorithms to teach a machine to perform specific tasks and learn independently from data (França et al., 2021). Data classification can be done using statistics such as mean, variance, and many more (Vijithananda et al., 2022). Using other statistics, such as cumulative summary, is also possible. Previous machine learning research includes using methods like Extra Trees and Support Vector Machine (SVM) to predict breast cancer risk factors (Alfian et al., 2022), AdaBoost to determine whether someone has alcohol use disorder (Park et al., 2021), Random Forest to predict loan (Sathish Kumar et al., 2022), and Random Forest to recognize drunk driver (Li et al., 2020). There are also common classification methods that can be performed without machine learning, such as Logistic Regression in various health fields, among others (Schober & Vetter, 2021).

The results of differences in dielectric due to evaporation over time of ethanol and methanol can be processed using machine learning with various conditions such as regularization and cross-validation on various classifiers. It can differentiate the type of alcohol (methanol or ethanol) tested based on the training model that has been carried out on machine learning. With the test results, it is hoped that the training model of a classifier will be able to distinguish whether the results of the evaporation trend over time in a new sample are ethanol or methanol compounds.

With only two classes, namely ethanol or methanol, as a binary classification, the focus of the research is to determine the potential of CSM in distinguishing types of alcohol (methanol or ethanol) from evaporation and data processing using machine learning, which is processed using Python. The research is limited to the two types of alcohol that are most widely circulated in society, namely ethanol and methanol. It can be further developed to determine other types of compounds and levels of other volatile solutions using different dielectrics.

Finally, the use of CSM to classify ethanol or methanol using machine learning based on data on differences in dielectric due to evaporation of the material has the potential to be a cheaper alternative to GC instruments. This method can serve as an initial alternative, especially for the bioethanol and biomethanol sectors, as well as small to medium-scale industries that require quick analysis. Determining levels using sensors and methods has never been done by any authors, leading to the novelty of the research.

II. METHODS

The data used are collected from an instrument for one hour with a five-second reading interval. Commercially procured 99% ethanol (Merck, 1.00983.2500) and 99% methanol (Merck, 1.06009.2500) serve as the test samples. Each reading is carried out 37 times for each type of sample (ethanol and methanol) with two CSM sensors for every measurement, which is then ready to be processed using machine learning in Python. The flow process diagram in the research is shown in Figure 1.

Figure 1 shows the flow process diagram in the research. The construction and measurement tool in the research is built using An Arduino UNO board with two CSM sensors (CSM 1 and CSM 2) as the main sensors to measure sample evaporation, Real Time Clock (RTC) DS3231SN to determine the measurement time, and BME 280 to determine environmental conditions in this case temperature, humidity, and pressure. The tool is placed in a closed container to minimize environmental interference and is connected to an adapter to read each datum automatically after the sample is dropped. There is a two-minute break to clean up the remaining solution, which has not evaporated for an hour, and drip the samples back into the two CSM plates. The tool used

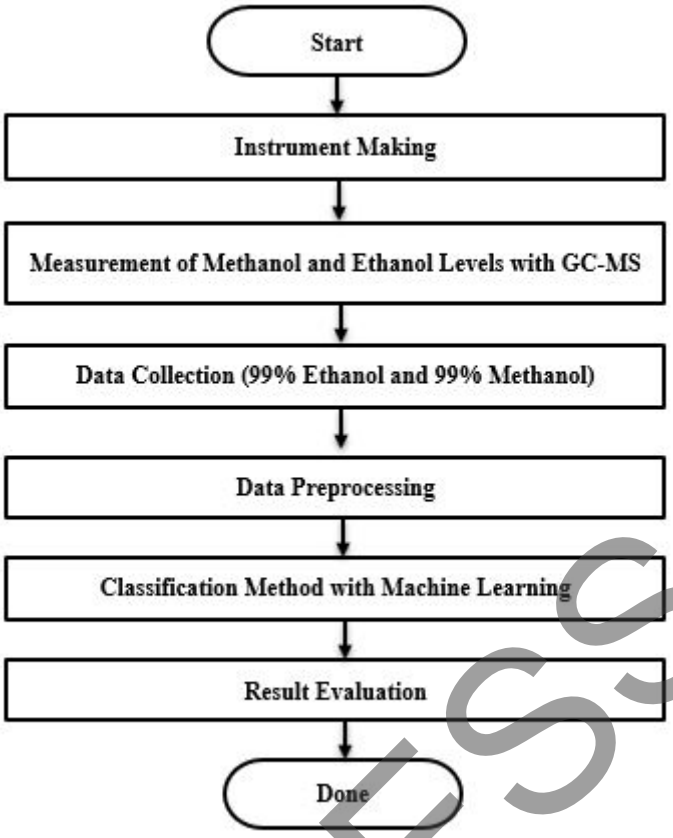


Figure 1 Flow Process Diagram in the Research

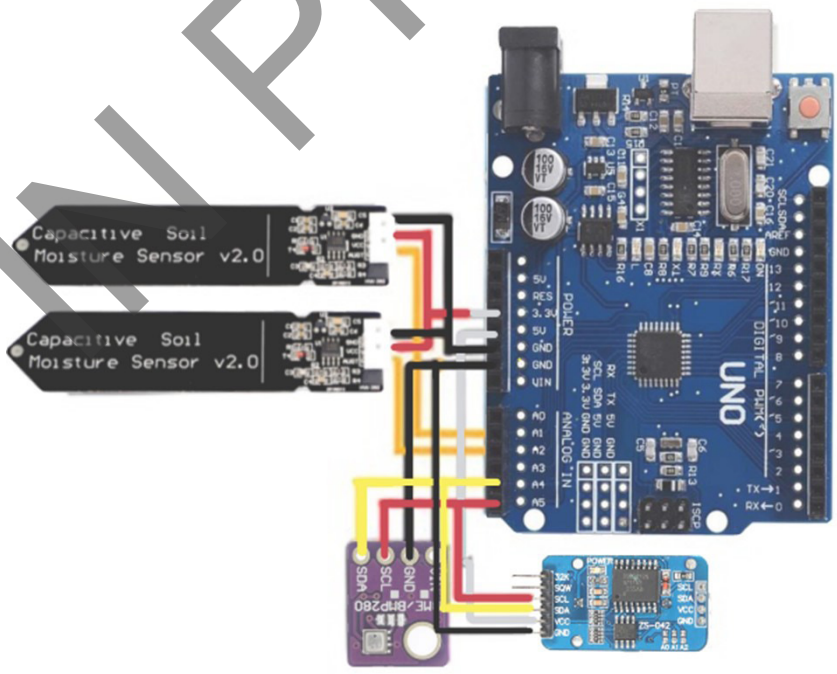


Figure 2 Capacitive Soil Moisture (CSM) Yield Graph for 99% Methanol and 99% Ethanol against Index (Time)

in the research is shown in Figure 2. Figure 2 shows that CSM 1 and 2 are respectively connected to pins A1 and A2 on the Arduino UNO.

Before checking using the instrument tool, the original methanol and ethanol levels are checked using GC-MS in a testing laboratory. The results of checking the levels are in the form of a wide response area of the chromatogram. Then, it is calculated in percentages to determine the levels.

Data collection using the instrument tool is carried out using a pipette to take each sample using a pipette filler and a pipette of 0.2 ml. The samples are 99% ethanol and 99% methanol. Each sample taken is then placed evenly on the CSM plate, with the solution leveling limit determined for each CSM.

The data collected are then preprocessed before being processed further. The data preprocessing on each measurement result is done by changing the time data in each txt file resulting from the tool for each series of evaporation data of a concentration in the *n*-th replication into date time form. Then, it is converted again into a timestamp. The timestamp is set so that each datum has a measurement time range in the form of an index in seconds from 0 to the *n*-th time, which is the same. The 1 hour is used for each sample with an interval of 5 seconds, so there are 720 data per sample. Then, initial data processing is carried out to remove outliers, with the outlier criteria being the results of CSM 1 and 2 sensor readings, which have values much higher or lower than several adjacent neighboring data with a certain threshold. There are many outliers at the beginning and end of the measurement, so only the 50th to 650th data out of 720 data (1-hour measurements at 5-second intervals) are taken for processing. Then, the data are smoothed using the Savitzky-Golay filter (savgol_filter) so that outliers can be minimized. Savgol_filter is used for smoothing data in the research because it is effective in reducing noise while preserving key features like peaks and curves, which are very important to maintain (Baihaqi et al., 2021).

The model is defined by classifying 99% ethanol and 99% methanol data based on the results of each CSM sensor and the environmental conditions read by Bosch Measurement Environmental (BME) in each file. Each data reading result for 1 hour is divided into several parts (n), and the cumulative total value is calculated for each part to differentiate 1 data point from other data points with similar characteristics based on the cumulative total. The goodness of measurement results of these features uses the accuracy of each classifier (32 classifiers) that is most optimal in each division (n). The change in element values used in each classifier for machine learning models is in the range of 20 to 50, so the most optimal element values can be determined to determine accuracy. Apart from the cumulative number, standard deviation, average, and difference parameters are also determined to become data classifier features. The research uses several formulas: Standard Deviation (SD) in Equation (1), the average ($\bar{x}$) in Equation (2) (Quirk & Palmer-Schuyler, 2020), the difference (D_i) in Equation (3), and the cumulative summary (*Cumsum*) in Equation (4) (Tygert, 2021).

$$SD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (1)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad (2)$$

$$D_i = x_{i+1} - x_i, \quad (3)$$

$$Cumsum = \sum_{i=1}^{n} x_i. \quad (4)$$

The formulas have n as the number of samples, x_i as a series of n sample values, and $\bar{x}$ as the average (Quirk & Palmer-Schuyler, 2020). Then, D_i is the difference between x elements at index $i + 1$ (Tygert, 2021). The Logistic Regression with cross-validation and various types of existing classifiers are used to determine the classification with the best accuracy. In determining just two compounds, namely 99% methanol and 99% ethanol, the probability is determined using the Sigmoid function shown in Equation (5) (Liu et al., 2021). It has β_0, β_1, ..., β_p as the regression coefficients β, x_1, x_2, ..., x_p as the input features of x, and e as the Euler number. The Sigmoid probability used in Logistic Regression calculates class $y = 1$. For an alternative class, namely $y = 0$, the probability is 1-$P(y = 1|x)$ (Liu et al., 2021).

$$P(y = 1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)}}. \quad (5)$$

In Logistic Regression, model training uses regularization and an appropriate solver. Solver is an algorithm to find optimal parameters for the Logistic Regression model, and regularization is a technique to prevent overfitting by adding a penalty term to the model's loss function (Nur et al., 2023). Additional tests are conducted using several solvers with $L1$, $L2$, and Elastic Net regularization according to the type of solver used. The $L1$, $L2$, and Elastic Net regularization formulas, together with the loss function in classification and logistic loss, are shown in Equations (6) to (9) (Nur et al., 2023).

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} y_i \log\left(h_\beta(x_i)\right) \\ +(1-y_i)\log\left(1-h_\beta(x_i)\right)\end{bmatrix}, \quad (6)$$

$$L1 = LogLoss + \lambda\sum_{j=1}^{p}|\beta_j|, \quad (7)$$

$$L2 = LogLoss + \lambda\sum_{j=1}^{p}\beta_j^2, \quad (8)$$

$$Elastic\ net = LogLoss + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2\right) \quad (9)$$

The *LogLoss* is logistic loss. It is the loss function of an optimized classification method, where

n is the number of samples, y_i is the actual label (0 or 1) for sample i, and $h_\beta(x_i)$ is the probability that has been predicted by the model for sample i by calculation using Sigmoid function. In Logistic Regression, the prediction results belonging to a certain class are calculated using a logistic function, which produces a value between 0 and 1 with a Sigmoid function for binary classification of Logistic Regression (Parhusip et al., 2020). Then, $L1$ is Lasso regularization, and $L2$ is Ridge regularization, where λ is a regularization parameter that controls the amount of penalty for both $L1$ and $L2$, β_j is the regression coefficient for the j feature, and in elastic net, α is a parameter that controls the division of the $L1$ and $L2$ penalty ratio and value between 0 and 1 (Nur et al., 2023).

In the Logistic Regression model, $L2$ regularization is used in Newton Cholersky because this solver is suitable for second-order derivatives for a smooth and differentiable penalty like $L2$, $L1$ regularization is used in Liblinear solver because Liblinear can handle non-smooth optimization problem form in $L1$, and Stochastic Average Gradient (Saga) solver is used for elastic net because this solver can optimize both penalty combination with variance reduction. Every regularization with every solver is calculated with the default λ parameter in Python, namely 1.0, and each model is used on 148 data (74 data from CSM 1 and 74 data from CSM 2) with 80% training data (118 data) and 30 testing data. Then, 60 features are used. They are adjusted to the best results from changes in the values of dividing elements in the data for standard deviation results, cumulative values, and averages that provide the best accuracy in the model.

The final step in the research is evaluation of the results. Evaluation of research results is carried out by determining the accuracy of each model in machine learning with various classifiers. In determining various types of classifiers, the accuracy is determined to be 90% or more for the type of classifier, which is considered to have good results.

In the evaluation of the optimal solver and penalty for the Logistic Regression model, performance is further assessed using a confusion matrix and the Area Under Curve (AUC) from the Receiver Operating Characteristic (ROC) curve. The ROC curve provides a graphical representation of the True Positive Rate (TPR) versus the False Positive Rate (FPR). The AUC, representing the area under the ROC curve, offers a measure of the model's sensitivity and specificity. An AUC value exceeding 0.8 is considered a strong model performance (Rizzo et al., 2023).

III. RESULTS AND DISCUSSIONS

The results of determining the purity levels of ethanol and methanol used as samples for research based on GC-MS are shown in Table 1. The GC-MS test results show that the original written content of pure ethanol and methanol as the material used does not have a content value reaching 99% but only 98.848% for ethanol and 96.071% for methanol.

Table 1 Sample Purity Results from Gas Chromatography-Mass Spectrometry (GC-MS)

No	Name	Retention Time	Content (%)
1.	Ethanol	2.415	98.848
2.	Methanol	1.911	96.071

The results of the tool reading are in the form of a txt file containing six columns, namely Timestamp, CSM 1 and CSM 2 for both CSM sensors, Temperature for temperature, Humidity for air humidity, and Pressure for air pressure with 720 lines of data. Each result from the CSM sensor, i.e., CSM 1 and 2, against time (Timestamp), is depicted using a graph in Python. The pattern of pure ethanol and methanol samples shows quite different results based on the evaporation graph produced (Figure 3).

The evaporation patterns observed in Figure 3 suggest the potential of CSM sensors to distinguish between the two sample types. This finding paves the way for the classification of various samples based on their evaporation characteristics. For each 99% ethanol and 99% methanol dataset, preprocessing is carried out in the form of calculating the SD, average, cumsum, and diff for each n row (2-50) of the 600 selected data, which are different from each other so that classification can be carried out using machine learning. In the initial (samples 1−49) and final segments (samples 651−720) of the dataset, numerous outliers are identified due to unstable evaporation measurements. Early in the process, fluctuations are significant as the system has not yet stabilized. In the final phase, values change drastically due to the minimal remaining liquid volume since most have evaporated.

Only stable data segments are utilized to avoid excessively high or low values, standardize the dataset, and mitigate the presence of outliers to address these issues. Additionally, the Savgol filter is applied as a smoothing technique. This filter effectively minimizes sudden spikes in the data caused by unstable electrical currents during measurement by averaging the fluctuations. The application of the Savgol filter ensures a more reliable dataset by softening abrupt changes and enhancing the overall stability of the recorded measurements.

In Logistic Regression classification, there are several provisions, namely cross-validation with seven folds (k) and several types of regularization and solvers to prevent overfitting (Table 2). The fold value setting is adjusted because too many fold values can cause overfitting even though they tend to be stable, and fold values that are too low can make the model performance inaccurate, so evaluation errors may occur (Chollet, 2021). In this case, the fold value is

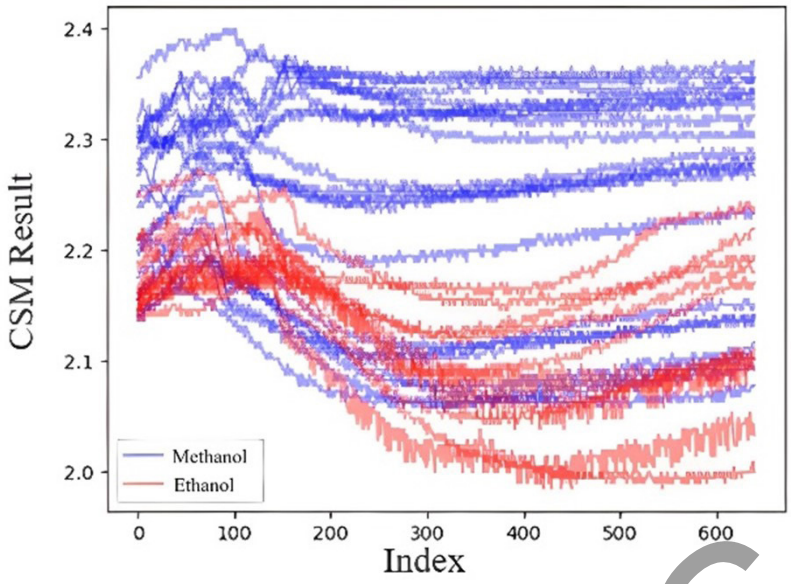


Figure 3 The Evaporation Patterns from Capacitive Soil Moisture (CSM) Data for 99% Methanol and 99% Ethanol against Index (Time)

Table 2 Accuration Results from Various Types of Regularizations and Solvers in Logistic Regression

No	Solver	Penalty	Accuracy (%)
1.	Newton-Cholesky	L2	93.33
2.	Liblinear	L1	90.00
3.	Saga	L1 : L2 = 25 : 75	86.67
		L1 : L2 = 50:50	86.67
		L1 : L2 = 75 : 25	86.67

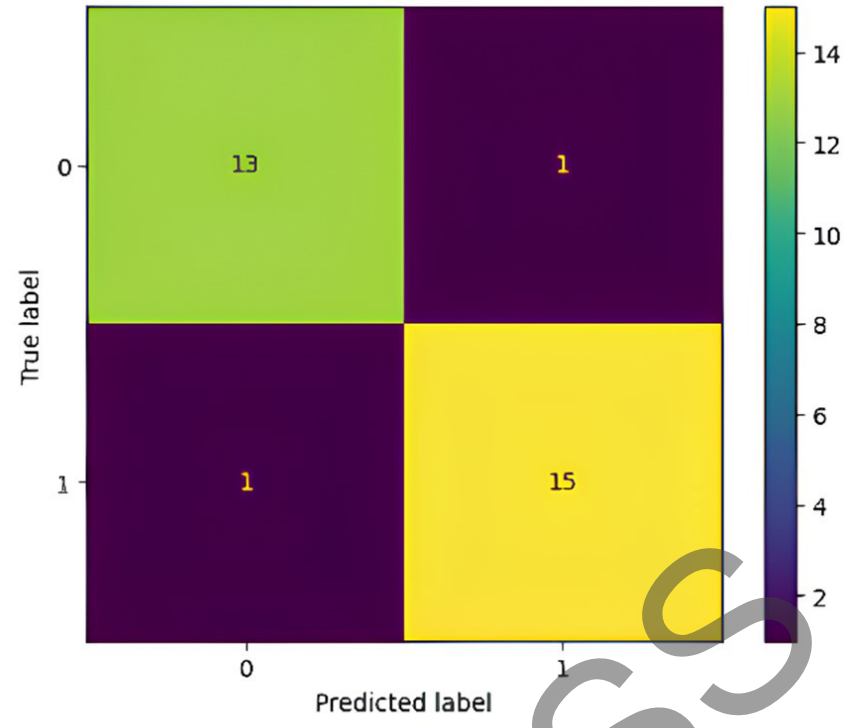
increased to seven to overcome data limitations and provide optimal model results, thereby reducing evaluation errors and preventing fold overfitting in *L2* regularization. The solver used is Newton Cholesky which is commonly used in Logistic Regression (Li, 2023). The Cholesky Newton solver works by minimizing a loss function that can only be paired with no penalty or *L2* penalty (Gupta et al., 2021). *L2* can handle limited data but has many features by squaring the coefficients to suit the research conditions.

Although no significant differences are observed in accuracy across various solvers and regularization techniques, the highest accuracy is achieved using the Newton-Cholesky solvers with *L2* regularization. From the accuracy results in Table 2, the Newton Cholesky solver provides accuracy with the highest results, 93.33%. It shows that Newton Cholesky with *L2* regularization provides good accuracy results and can model the data well. The Liblinear solver with *L*1 regularization also has good accuracy, up to 90.00%. Liblinear is a solver that is suitable for large datasets. However, because the data used are limited, this solver is less suitable for use. Further evaluation using the Confusion Matrix (Figure 4) and ROC results (Figure 5) is carried out for the best Logistic Regression model. It uses Logistic Regression with the Newton-Cholesky solver and *L*2 penalty.

Figure 4 shows the confusion matrix of the best Logistic Regression model with Newton-Cholesky solver and *L*2 penalty. This model demonstrates balanced classification, with True Positives (15), True Negatives (13), and Precision and Recall of 0.93 for class 0 and 0.94 for class 1. It leads to consistent F1 scores for both classes. This result highlights the model's effectiveness in distinguishing ethanol and methanol based on dielectric and evaporation rate differences.

The AUC from the ROC curve is 0.92 in Figure 5. An AUC value of more than 0.8 is considered to indicate that the model provides high accuracy (Rizzo et al., 2023). So, it is determined that the Logistic Regression model with an AUC of 0.92 can identify positive and negative instances well.

Further evaluation using 32 different classifiers in Python, without additional data processing, reveals that the Extra Trees, Random Forest, and Gaussian Naive Bayes achieve the highest accuracy of 96.67% on testing data. The AdaBoost Classifier closely



	precision	recall	f1-score	support
0	0.93	0.93	0.93	14
1	0.94	0.94	0.94	16
accuracy			0.93	30
macro avg	0.93	0.93	0.93	30
weighted avg	0.93	0.93	0.93	30

Figure 4 Confusion Matrix of Logistic Regression with Newton-Cholesky Solver and *L*2 Penalty

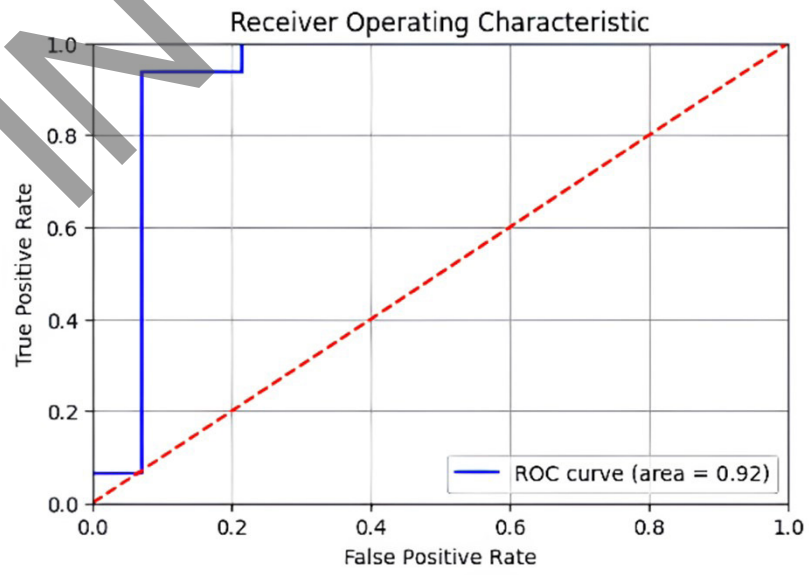


Figure 5 Receiver Operating Characteristic (ROC) in Logistic Regression with Newton-Cholesky Solver and *L*2 Penalty

Table 3 Classifier Accuracy with Change in Element Values (20 to 50)

No	Classifier	Accuracy (%) in Every Element Values			
		20	30	40	50
1.	Extra Trees Classifier	96.67	96.67	96.67	96.67
2.	Random Forest Classifier	96.67	96.67	96.67	96.67
3.	AdaBoost Classifier	93.33	93.33	93.33	93.33
4.	Gaussian Naive Bayes	96.67	96.67	96.67	96.67
5.	Logistic Regression	86.67	86.67	86.67	86.67
6.	Gradient Boosting Classifier	90.00	90.00	90.00	90.00
7.	Ridge Classifier Cross Validation	90.00	90.00	90.00	90.00

follows, with 93.33% accuracy. Changes are made to element values to determine evaluation based on the differences in accuracy. Only classifiers with an accuracy of more than 90% of 32 classifiers in Python are selected, which are considered good (Table 3).

Based on the results of the accuracy of changing the elements taken, it is determined that changing the number does not significantly affect the accuracy results. It still produces the three best models: Extra Trees, Random Forest, and Gaussian Naive Bayes. However, it can be determined that features with too few elements will make the model too complex, while those that are too large can be unrepresentative of the data. Hence, it is determined that an element value of 50, even if it is large, is enough to provide the best accuracy in each classifier naturally without changing the components in the classifier.

Changes in accuracy can occur as a result of the data processing process in the form of changing the array using the selection of certain elements. The array transformation resulting from changes in the values of selected elements greatly influences the follow-up process. The greater the value of the variable is, the less data will be processed. So, it tends to be more general to data fluctuating, and there will be changes in data patterns.

The changes in accuracy in some classifiers are not statistically significant. However, the results suggest that using certain conditions, such as selecting every 50^{th} element in an array and calculating the difference between consecutive elements, introduces more precise modifications to the data. These methods effectively alter the representation of values within the input array without changing the original data (Harris et al., 2020). In addition, since classification is carried out without overfitting prevention, such as cross-validation or regularization for relatively limited data with many features, there is still the possibility of the model overfitting to very good accuracy. In addition, no evaluation of ROC, AUC, or confusion matrix is carried out on the 32 classifiers because there is a tendency for overfitting. Then, there is no optimization of the parameters of each classifier used, and more focus is placed on several elements that can be determined from the measurements. However, the accuracy results of each classifier indicate that the data can be predicted well using the models even without any preprocessing techniques.

Based on the results, it is determined that the machine learning method using Random Forest, Extra Trees, Logistic Regression with cross-validation with seven folds, *L*2 regularization, Newton-Cholesky and Liblinear solver, Gaussian Naive Bayes, and AdaBoost on the CSM reading results of both types of alcohol (ethanol and methanol) provides very good accuracy, above 90%. The values are 96.67%, 96.67%, 93.33%, 96.67%, and 93.33% for Extra Trees, Random Forest, Logistic Regression with several conditions, Gaussian Naive Bayes, and AdaBoost, respectively. These five types of classifiers are determined to differentiate types of alcohol based on evaporation. It is due to evaporation speed and changes in numbers related to the dielectric of the material accurately. High accuracy shows that most data are predicted correctly based on the training results using machine learning.

The results indicate that the tool effectively distinguishes between types of alcohol based on the distinct evaporation characteristics and dielectric properties of each sample type. Although the research is limited to two alcohols (ethanol and methanol), this finding suggests that CSM has the potential for differentiating other volatile compounds, particularly those with low boiling points, such as various other alcohols, based on their unique dielectric properties.

Though not yet implemented physically, the classification results from this tool can be extended with an LCD display that shows classification predictions, such as "ethanol" or "methanol". This setup can further evolve into a capable device of determining ethanol-methanol levels in a mixture using regression methods. The proposed monitoring system can enable rapid and cost-effective identification of ethanol or methanol in raw materials or final products, providing a feasible alternative to conventional methods. However, it is currently tested only on simulated data.

A potential implementation scenario for this tool is in the bioethanol or biomethanol industries, where it can be used to identify pure ethanol and methanol without contaminants for potential use in gasoline. The usage process will proceed as follows:

a sample is placed on the CSM sensor, which records the evaporation profile over one hour to capture the distinctive evaporation pattern of ethanol or methanol. The recorded data are then analyzed by machine learning models stored in the device's memory like Arduino UNO, with the predicted result displayed on an LCD screen as "ethanol" or "methanol".

IV. CONCLUSIONS

In the research, the classification of alcohols, specifically ethanol and methanol, as a cost-effective alternative to GC has been successfully demonstrated using a method that determines evaporation rates based on dielectric differences via CSM and machine learning classification. Specifically, the difference in the dielectric of the evaporation speed of ethanol and methanol read by CSM can be used to differentiate between pure ethanol and methanol by classification using several machine learning methods. The logistic regression with cross-validation with seven folds, Newton Cholesky solver, and *L*2 regularization with an accuracy of 93.33% and an AUC value of the ROC curve of 0.92. It indicates that the model provides good prediction accuracy for ethanol and methanol. Furthermore, without further processing for hyperparameters, the Random Forest classifier, Extra Trees Classifier, and Gaussian Naive Bayes can differentiate ethanol and methanol with an accuracy of 96.67%, followed by the AdaBoost Classifier with an accuracy of 93.33%. It is the same as logistic regression with various regularization and cross-validation conditions.

However, the research is limited to two types of commonly used alcohols with a small dataset, but it serves as a preliminary step for future research involving various volatile solutions. Additionally, the machine learning classification is conducted without hyperparameter optimization. However, it demonstrates that the research can be applied to different machine learning algorithms with reasonably good accuracy. The research can also be further developed for other types of alcohols, and the concentration proportion between two volatile liquids can be determined by regression. Therefore, future research can be explored in a different solution with different dielectric evaporations.

REFERENCES

Alfian, G., Syafrudin, M., Fahrurrozi, I., Fitriyani, N. L., Atmaji, F. T. D., Widodo, T., ... & Rhee, J. (2022). Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. *Computers*, *11*(9), 1–14. https://doi.org/10.3390/computers11090136

Baihaqi, M. Y., Lumoindong, C. W. D., & Vincent. (2021). Simulasi perbandingan filter Savitzky Golay dan filter Low Pass Butterworth pada orde ketiga sebagai pembatal kebisingan. *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, *1*(2), 226–232. https://doi.org/10.24002/konstelasi.v1i2.4294

Chollet, F. (2021). *Deep learning with Python, second edition*. Manning Publications.

França, R. P., Monteiro, A. C. B., Arthur, R., & Iano, Y. (2021). An overview of deep learning in big data, image, and signal processing in the modern digital age. *Trends in Deep Learning Methodologies: Algorithms, Applications, and Systems*, 63–87. https://doi.org/10.1016/B978-0-12-822226-3.00003-9

Gupta, A., Parmar, R., Suri, P., & Kumar, R. (2021). Determining accuracy rate of artificial intelligence models using Python and R-Studio. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 889–894). https://doi.org/10.1109/ICAC3N53548.2021.9725687

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hrisko, J. (2020, July 5). *Capacitive soil moisture sensor theory, calibration, and testing*. https://makersportal.com/s/capacitive_soil_moisture_sensors_joshua_hrisko.pdf

Iliev, S. (2021). A comparison of ethanol, methanol, and butanol blending with gasoline and its effect on engine performance and emissions using engine simulation. *Processes, 9*(8), 1–14. https://doi.org/10.3390/pr9081322

Joseph, J. A., Akkermans, S., & Van Impe, J. F. (2022). Processing method for the quantification of methanol and ethanol from bioreactor samples using gas chromatography–Flame ionization detection. *ACS Omega, 7*, 24121–24133. https://doi.org/10.1021/acsomega.2c00055

Korban, A., Charapitsa, S., Čabala, R., Sobolenko, L., Egorov, V., & Sytova, S. (2021). Advanced GC–MS method for quality and safety control of alcoholic products. *Food Chemistry, 338*. https://doi.org/10.1016/J.FOODCHEM.2020.128107

Li, M. (2023). Teaching business analytics students logistic regression using Python and R. *Business Education Innovation Journal, 15*(1), 35–41.

Li, Z., Wang, H., Zhang, Y., & Zhao, X. (2020). Random forest–based feature selection and detection method for drunk driving recognition. *International Journal of Distributed Sensor Networks, 16*(2).

Liu, Y., Esan, O. C., Pan, Z., & An, L. (2021). Machine learning for advanced energy materials. *Energy and AI, 3,* 1–27. https://doi.org/10.1016/J.EGYAI.2021.100049

Lone, B. G., Undre, P. B., Patil, S. S., Khirade, P. W., & Mehrotra, S. C. (2008). Dielectric study of methanol–ethanol mixtures using TDR method. *Journal of Molecular Liquids, 141*(1–2), 47–53. https://doi.org/10.1016/j.molliq.2008.03.001

Mohsen-Nia, M., Amiri, H., & Jazi, B. (2010). Dielectric constants of water, methanol, ethanol, butanol and acetone: Measurement and computational study.

Journal of Solution Chemistry, 39, 701–708. https://doi.org/10.1007/s10953-010-9538-5

Nur, A. R., Jaya, A. K., & Siswanto. (2023). Comparative analysis of ridge, LASSO, and elastic net regularization approaches in handling multicollinearity for infant mortality data in South Sulawesi. *Jurnal Matematika, Statistika dan Komputasi, 20*(2), 311–319. https://doi.org/10.20956/j.v20i2.31632

Paolini, M., Tonidandel, L., & Larcher, R. (2022). Development, validation and application of a fast GC-FID method for the analysis of volatile compounds in spirit drinks and wine. *Food Control, 136*. https://doi.org/10.1016/J.FOODCONT.2022.108873

Parhusip, H. A., Susanto, B., Linawati, L., Trihandaru, S., Sardjono, Y., & Mugirahayu , A. S. (2020). Classification breast cancer revisited with machine learning. *International Journal of Data Science, 1*(1), 42–50. https://doi.org/10.18517/ijods.1.1.42-50.2020

Park, S. J., Lee, S. J., Kim, H., Kim, J. K., Chun, J. W., Lee, S. J., ... & Choi, I. Y. (2021). Machine learning prediction of dropping out of outpatients with alcohol use disorders. *PLoS ONE*, *16*(8), 1–13. https://doi.org/10.1371/journal.pone.0255626

Putri, A., & Kasli, E. (2017). Pengaruh suhu terhadap viskositas minyak goreng. In *Prosiding Seminar Nasional MIPA III* (pp. 464–469).

Quirk, T. J., & Palmer-Schuyler, J. (2020). Sample size, mean, standard deviation, and standard error of the mean. In *Excel 2019 for human resource management statistics: A guide to solving practical problems*. Springer. https://doi.org/10.1007/978-3-030-58001-8_1

Rizzo, V., Salmasi, M. Y., Sabetai, M., Primus, C., Sandoe, J., Lewis, M., ... & Athanasiou, T. (2023). Infective endocarditis: Do we have an effective risk score model? A systematic review. *Frontiers in Cardiovascular Medicine*, 10, 1–12. https://doi.org/10.3389/fcvm.2023.1093363

Sathish Kumar, L., Pandimurugan, V., Usha, D., Guptha, M. N., & Hema, M. S. (2022). Random forest tree classification algorithm for predicating loan. *Materials Today: Proceedings, 57*, 2216–2222. https://doi.org/10.1016/J.MATPR.2021.12.322

Savchuk, S. A., Palacio, C., Gil, A., Tagliaro, F., Kuznetsov, R. M., Brito, A., & Appolonova, S. A. (2020). Determination of the chemical composition of alcoholic beverages by gas chromatography-mass spectrometry. *Journal of Food Processing and Preservation, 44*(9). https://doi.org/10.1111/jfpp.14676

Schober, P., & Vetter, T. R. (2021). Logistic regression in medical research. *Anesthesia & Analgesia, 132*(2), 365–366. https://doi.org/10.1213/ANE.0000000000005247

Septiana, A. T., & Asnani, A. (2013). Antioxidan activity of Sargassum duplicatum seaweed extract. *Jurnal Teknologi Pertanian, 14*(2), 79–86.

Tygert, M. (2021). A graphical method of cumulative differences between two subpopulations. *Journal of Big Data, 8,* 1–29. https://doi.org/10.1186/s40537-021-00540-9

Vijithananda, S. M., Jayatilake, M. L., Hewavithana, B., Gonçalves, T., Rato, L. M., Weerakoon, B. S., ... & Dissanayake, K. D. (2022). Feature extraction from MRI ADC images for brain tumor classification using machine learning techniques. *BioMedical Engineering Online, 21*(1), 1–21. https://doi.org/10.1186/s12938-022-01022-6

Yanti, A., Mursiti, S., Widiarti, N., Nurcahyo, B., & Alauhdin, M. (2019). Optimalisasi metode penentuan kadar etanol dan metanol pada minuman keras oplosan menggunakan Kromatografi Gas (KG). *Indonesian Journal of Chemical Science, 8*(1), 53–59.