Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 2 (2023) 226 - 235 ISSN Media Elektronik: 2580-0760

Improving the Accuracy of C4.5 Algorithm with Chi-Square Method on Pure Tea Classification Using Electronic Nose

Mula Agung Barata[1], Edi Noersasongko[2], Purwanto[3], Moch Arief Soeleman[4]
[1] Teknik Informatika, Sains dan Teknologi, Universitas Nahdlatul Ulama Sunan Giri
[2,3,4]Magister Teknik Informatika, Ilmu Komputer, Universitas Dian Nuswantoro
[1]mula.ab26@gmail.com, [2]edi.noersasongko@dsn.dinus.ac.id, [3]purwanto@dsn.dinus.ac.id, [4]arief22208@gmail.com

Abstract

Tea is one of the plantation products within the Ministry of Agriculture of the Republic of Indonesia, which plays an essential role as a mainstay commodity that boosts the Indonesian economy. Each type of tea has different properties, and the aroma of each type of tea can measure the quality of the tea. The human sense of smell is still very limited in classifying pure types of tea. Therefore, a device is needed to help measure the aroma of tea from an electronic nose. The devices attached to several gas sensors help humans take data from the smell of pure tea and calculate the value of each type of tea to test datasets with data mining algorithms. This study uses the C4.5 algorithm as a classification method with advantages over noise data, missing values, and handling variables with discrete and continuous types. Meanwhile, Chi-square is used to perform attribute severing in the data preprocessing process to increase the accuracy of dataset testing. Testing a pure tea dataset with four whole attributes, namely CO_2, CO, H_2, and CH_4, using the C4.5 algorithm resulted in an accuracy of 93.65% and an increase in the accuracy performance of the C4.5 algorithm by 94.27% with dataset testing using Chi-Square feature selection with the two highest value attributes.

Keywords: electronic nose; e-nose; C4.5 algorithm; chi-square; tea plantation commodities; pure indonesian tea

1. Introduction

Tea is one of the plantation products within the Ministry of Agriculture of the Republic of Indonesia, which plays an essential role as a mainstay commodity that boosts the Indonesian economy [1]. In 2018, the Central Statistics Agency noted that tea could increase foreign export exchange by 1.5% of the Gross Domestic Product of the agricultural sector or by 108.5 million USD [2]. Each type of tea has different properties, and the aroma of each type of tea can measure the quality of the tea. Nowadays, some business people modify many types of tea by adding fragrance to the tea, which reduces the purity of the tea and the properties contained. The human sense of smell is still very limited in classifying pure tea types, so it isn't easy to distinguish which types of black, green, oolong or white tea [3]. Therefore, a method is needed to help retrieve data and distinguish tea's aroma with an instrumentation tool in the form of an electronic nose. Research on the application of the electronic nose by Jun Wang et al. in the evaluation of tea quality combined with chemometric methods this study concluded that the electronic nose shows the feasibility of its application in carrying out its classification with several methods used [4]. Further research on the application of electronic nose in the classification of civet coffee and not civet Indonesia based on a comparison of statistical analysis by Sulaiman Wakhid this study resulted in an accuracy of 97% of the Decision Tree algorithm testing and standard deviation statistical parameters [5].

Classification is one of the data mining algorithms that have the concept of grouping data into specific criteria by reading previously existing data. The concept of a classification algorithm is to predict the categorical class label of data to classify it into one of the specified classes [6]. Algorithm C4.5 is one of the classification algorithms that produce decision trees developed by Quinlan from the development of the ID3 algorithm as the previous generation algorithm, which is tested to group datasets with specific criteria by forming decision trees and has advantages over data noise, missing values, and handling variables with discrete and continuous types [7]. According to Burak F. Tanyu, his research proved that the C4.5 algorithm in testing balanced and unbalanced datasets produced

Accepted: 17-11-2022 | Received in revised: 08-02-2023 | Published: 03-03-2023

the highest percentage of accuracy, which was 99% compared to the C5.0 and Random Forest algorithms [8]. Meanwhile, Xiangfei et al. proved that the C4.5 algorithm could better handle datasets with continuous attributes in this study is an online electrical voltage stability assessment dataset [9]. Gite et al., in their research on machine learning-based intrusion detection for various types of attacks on the Wireless Sensor Network, proved that the C4.5 algorithm produces higher accuracy than CART [10]. A different research topic conducted by Sundaramurthy et al. tested the C4.5 algorithm in dealing with the Rheumatoid Arthritis disease dataset with a predicted percentage value of 86.36% [11].

The Chi-square method can remove unnecessary features and only use components that meet the threshold limit on predefined parameters [12]. In their research, said Discussing et al., the Chi-Square method proved the best F-measures with a percentage value of 90.50% in selecting attributes to improve the classification of Arabic texts [13]. Meanwhile, in their research, Nuran Peker et al. proved that the discrete data method handled by Chi-square achieved maximum results compared to the original data in its implementation for the merger of the classification techniques [14]. According to Chena et al. in their research on the topic of denoising optical coherence tomography images based on the similarity of Chi-square and Fuzzy Logic with satisfactory results with indicators of obtaining better visuals [15]. Meanwhile, Mahana et al., in their research, proved that the development of MFlexDT to Chi-MFlexDT is better in terms of accuracy and kappa value [16].

In addition to research on the application of electronic noses to detect tea vapor and the application of the C4.5 algorithm, researchers also reviewed research with the Chi-square method with the topic of conducting a feature selection process on the dataset used in the study. Research on applying the Chi-square method to combine classification algorithms by Nuran Peker and Cemalettin Kubat mentioned that there is often testing of the Chi-square method with classification algorithms such as Decision Tree and Naïve Bayes [14]. The paper in this research also mentioned that the research used a 10-fold cross-validation method testing to test the classification accuracy in this study.

Research on the electronic nose has been found in recent years. However, research on the classification of pure tea types that utilize the application of electronic noses with machine learning algorithm processing has yet to be found. Several studies by researchers show that the C4.5 algorithm is a reliable method because it can produce research with the highest predictive accuracy of other classification algorithms. In addition to research on the application of electronic noses for tea vapor measurement and the application of the C4.5 algorithm referred to researchers also reviewed research using the Chi-square method with the topic of conducting a feature selection process on the dataset used in the study. Based on this background, researchers used the C4.5 algorithm with the Chi-Square method to select features in a pure tea dataset. The Chi-square method is proposed to optimize the performance of the C4.5 algorithm in this study to obtain the best results in conducting pure tea dataset classification testing.

2. Research Method

2.1. Electronic Nose Design

The researcher took the object of study using a device developed by themselves and adapted to the object of study by taking into account various references to the development of previous electronic noses with different things [17-20].

An artificial or electronic nose is a fake nose that humans deliberately design with the principle of cooperation with the biological olfactory system. The device is designed to detect and recognize odors presented in the gas sensor circuit. These devices can be developed and applied in various foods and beverages to maintain consistency in the product's aroma [17]. The electronic nose device used to retrieve pure tea data in this study was compiled by several components, including the Arduino Uno AT Mega 256 microcontroller with several MQ Series sensors [5]. The gas sensor used in the design of the electronic nose can be seen in Table 1.

Table 1. Gas Sensors

Sensors	Detected Gases
MQ-135	Carbon dioxide (CO_2) Carbon monoxide (CO)
MQ-8	Hydrogen (H_2)
MQ-4	Methane (CH_4)

The gas sensor installed on the electronic nose device recognizes the gas entering the air circulation chamber. This air circulation chamber serves to break down the aroma of pure tea so that its value can be measured. The configuration scheme in the design of the electronic nose can be seen in Figure 1.

The design of electronic nose devices in this study results from a reference to the development of electronic nose devices in previous studies with different functions. Several studies related to the application of electronic nose state that the electronic nose functions as a coffee quality meter, rice quality meter and comparison of the difference between beef and pork. The results of the electronic nose device design in this study can be seen in Figure 2.

DOI: https://doi.org/10.29207/resti.v7i2.4687

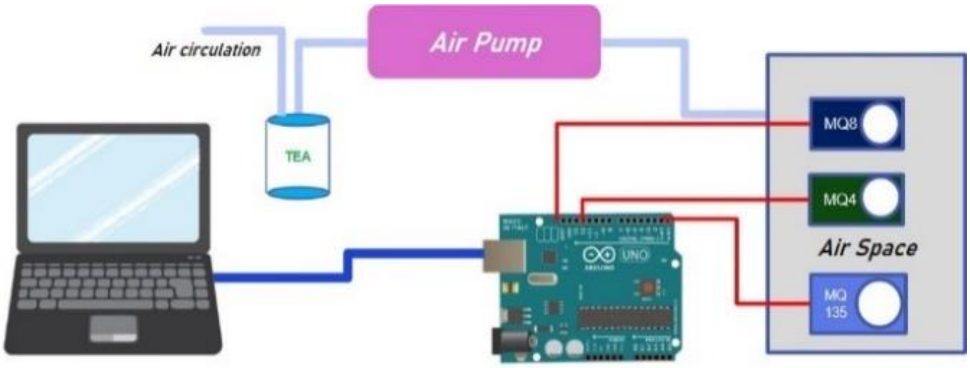


Figure 1. Electronic Nose Configuration Scheme

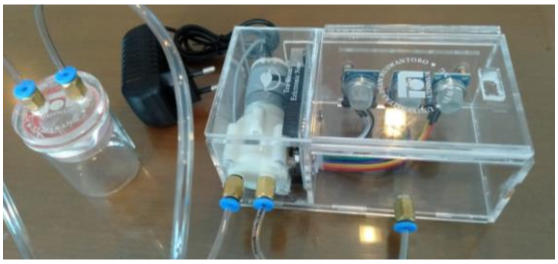


Figure 2. Electronic Nose Device

2.2. Device Communication System

The flow of the system designed on the electronic nose device, which involves communication between hardware and software to obtain a pure tea dataset, is shown in Figure 3. The four types of pure tea, namely black tea, green tea, oolong tea, and white tea, are measured in a specific size and dissolved in water, which boils at 100 degrees Celsius and then neutralizes the temperature to 25 degrees Celsius. After the water temperature has dropped to 25 degrees Celsius, the steam from the tea solution in the measuring cup begins to flow into the air chamber assisted by an air pump.

The flow of pure tea vapor from the measuring cup to the air chamber that has been installed gas sensors is measured for the concentration value in pure tea steam in units of PPM (Parts Per Million), namely CO_2, CO, H_2, and CH_4. Collecting data from tea steaming lasts for 5 minutes for each type of tea. The data collected from the concentration values in the four types of tea vapor were taken by three gas sensors installed with control from Arduino Uno, which was connected to a computer as a monitoring device during the data collection process. The data is stored in the computer, and then the data visualization process is carried out to find out the condition of the data obtained. The obtained dataset is then visualized before going through the pre-processing process. Data preprocessing is one of the processes that is carried out before the dataset is tested into an algorithmic model.

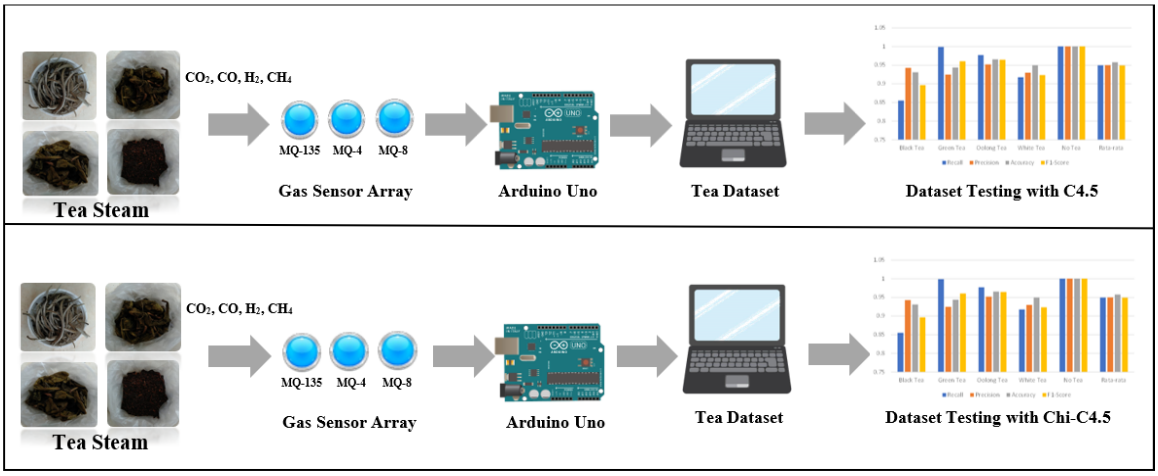


Figure 3. Device Communication System of Electronic Nose

2.3. Dataset

The object of this study is a dataset of four types of pure tea processed from the Camellia Sinensis leaf plant taken and processed by the Jamus Tea Garden under the management of PT. Candi Loka, located in the Ngawi Regency, East Java, is in the Girikerto Village Area, Sine District. The pure tea used in this study was derived from black tea, green tea, oolong tea and white tea.

The dataset (Table 2) in this study consists of four attributes based on the results of gas chromatography and mass spectrometry analysis of pure tea solution. The four data attributes are Carbon dioxide (CO_2), Carbon monoxide (CO), Hydrogen (H_2) and Methane (CH_4). The pure tea dataset used in this study consisted of 2756 data records from four types of pure tea native to Indonesia and four data attributes in each class.

Table 2. Pure Tea Dataset

No.	CO_2	CO	H_2	CH_4
1	302	428	25	67
2	302	428	25	67
3	302	428	25	67
4	302	428	25	67
5	302	428	25	67
6	302	428	25	67
7	302	428	26	67
8	302	428	26	67
9	302	428	25	67
10	302	428	26	67
11	302	428	26	67
12	302	428	26	67
13	302	428	26	67
14	302	438	26	68
15	302	428	26	67
16	302	438	26	68
17	302	438	25	68
…	…	…	…	…
2751	324	471	27	72
2752	324	471	27	72
2753	324	471	27	71
2754	324	471	27	72
2755	324	471	27	72
2756	324	471	27	72

2.4. C4.5 Algorithm

The C4.5 algorithm is a development of the ID3 algorithm, which is still a family with a Decision Tree. C4.5 is a classification model shaped like an inverted tree where C4.5 is easy for ordinary people to understand. C4.5 can work on categorical or numerical type attributes, overcome missing values, perform pruning processes and perform better than its predecessor methods, such as CART and ID3[11]. The formula formulated in calculating the C4.5 algorithm is in testing a dataset. The procedure for the calculation of the Entropy value is in Formula 1.

$$Entropy(S_1, S_2, \ldots, S_n) = \sum_{i=1}^{n} Pi * log_2(Pi) \quad (1)$$

S is the set of cases, n is thetotal of samples, and Pi is the class proportions.

While Formula 2 used in calculating the Gain value is as follows:

$$Gain(A) = Entropy(S) - \sum_{i=1}^{n} \left(\frac{|Si|}{|S|}\right) * Entropy(Si) \quad (2)$$

S is the set of cases, A is the attributes. N is the total of samples, |Si| is the number of cases on the partition, and |S| the total of cases in S.

Then calculate the value of Split Info with Formula 3.

$$Split\ Info(S, A) = \sum_{i=1}^{0} \left(\frac{|St|}{|S|}\right) * log_2\left(\frac{|St|}{|S|}\right) \quad (3)$$

S is the set of cases, A is the attributes, and |Si| is the number of cases on the partition,

After the Gain and Split Info values are found, then calculate the Gain Ratio value with Formula 4.

$$Gain\ Ratio(S, A) = \frac{Gain\ Ratio(S,A)}{Split\ Information(S,A)} \quad (4)$$

S is the set of cases, and A is the attributes.

2.5. Chi-Square

Chi-square is one of the statistical methods used in testing the significance of the relationship between the value of a variable and a class based on the level of relevance to express the similarity of adjacent intervals [21]. Formula 5 is the calculation of Chi-square.

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (5)$$

x^2 is the chi-square, O_i is the observed value and E_i is the expectation value

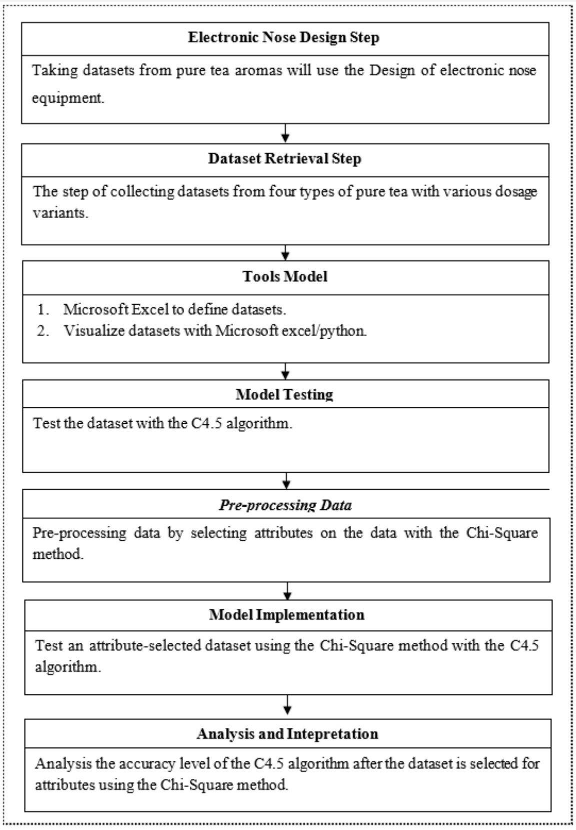
2.6. Research Framework

The stages in this study are outlined in a frame of mind based on the sequence of steps that the researcher performs during the survey. The framework of thought is seen in Figure 4.

The first step in this study is to design an electronic nose device to retrieve datasets as objects in this study. The design includes selecting hardware according to the needs and research references related to similar devices, device configuration, and sensor calibration testing in the machine.

The second step is taking datasets from four types of pure tea vapors that have been measured in a specific size. Data collection with electronic nose devices with several provisions that refer to a particular time and length is calculated based on related research references.

The third step is to define and visualize the dataset of the four types of pure tea vapor into a model with a tool in the form of excel or python so that the condition of the dataset obtained can be understood to get ease in handling the dataset in this study before being tested into the C4.5 algorithm model.



Electronic Nose Design Step

Taking datasets from pure tea aromas will use the Design of electronic nose equipment.

Dataset Retrieval Step

The step of collecting datasets from four types of pure tea with various dosage variants.

Tools Model

1. Microsoft Excel to define datasets.
2. Visualize datasets with Microsoft excel/python.

Model Testing

Test the dataset with the C4.5 algorithm.

Pre-processing Data

Pre-processing data by selecting attributes on the data with the Chi-Square method.

Model Implementation

Test an attribute-selected dataset using the Chi-Square method with the C4.5 algorithm.

Analysis and Intepretation

Analysis the accuracy level of the C4.5 algorithm after the dataset is selected for attributes using the Chi-Square method.

Figure 4. Research Framework

The fourth step is the stage in the trial of the algorithm model used in this study to classify the dataset obtained. This study used the C4.5 algorithm model to conduct classification testing of datasets to get accurate test results.

In the fifth step, researchers strive to improve the accuracy obtained from testing the C4.5 algorithm on datasets by conducting a data preprocessing process with the features selection method. The feature selection method used is Chi-square, believed to help improve the accuracy of the C4.5 algorithm.

The sixth step is the model implementation stage by retesting datasets that have gone through a preprocessing process with the Chi-square method to determine the percentage of the accuracy improvement of the C4.5 algorithm testing of pure tea datasets.

The last step is conducting analysis and interpretation of the test results in this study and knowing the comparison of the level of accuracy produced between testing the C4.5 algorithm against a pure tea dataset with a dataset that has gone through the preprocessing process and then tested with the C4.5 algorithm.

At this analysis and interpretation stage, the researcher concludes the research results. The results of this study will be outlined in the conclusion section of the survey. In addition to conclusions, suggestions from researchers are also submitted in the same area for further research improvement in terms of research methods and models of algorithms and objects of research. So that further research obtains improvements and produces a higher level of accuracy than the measurement results using the selected model.

2.7. Proposed Method

Researchers chose the method in this study to classify the types of pure tea to assist refined tea producers in classification the types of tea to be produced according to the properties contained in each kind of tea. Based on the proposed method can be presented through the method scheme in Figure 5 in a fundamental overview of the flow of testing datasets with the C4.5 algorithm.

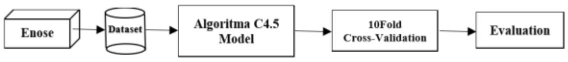


Figure 5. Testing Datasets with C4.5 Algorithm

Based on the proposed method can be presented through the method scheme in Figure 6 in an accurate picture of the flow of testing the dataset through the attribute selection process with the Chi-square method and then experimenting with the C4.5 algorithm.

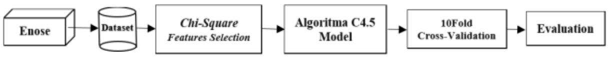


Figure 6. Testing Datasets with Chi-Square and C4.5

2.8. Evaluation

The evaluation of experiments conducted in this study uses a confusion matrix to measure the level of correctness of the performance of the C4.5 algorithm in carrying out the classification process on pure Indonesian tea datasets. Four classes matrix was used in this study because it adjusted the number of classes of datasets processed. The matrix pattern in this study is shown in Table 3.

Table 3. Confusion Matrix 4 Class

		Predictive			
	Class	A	B	C	D
Actual	A	TP	FP	FP	FP
	B	FN	TP	TN	TN
	C	FN	TN	TP	TN
	D	FN	TN	TN	TP

From the matrix modelling in table 3, recall, precision, accuracy and F-1 Score values can be calculated. The formula for each of these calculations is in Formula (6), (7), (8) and (9).

$$Recall = \frac{TP}{TP+FN} \qquad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Accuracy = \frac{TP+TN}{(TP+FP)+(FN+TN)} \quad (8)$$

$$F1\text{-}Score = 2.\frac{Recall\,.\,Precision}{Recall+\,Precision} \quad (9)$$

TP is true positives, FP is false positives, TN is true negatives, and FN is false negatives.

3. Results and Discussions

The test in this study begins with data preparation, which will later test with two experimental scenarios. The first experiment tested the dataset with the C4.5 algorithm and the second experiment tested the dataset with a pre-processing process by selecting attributes using Chi-square before testing with the C4.5 algorithm to obtain higher accuracy results. Before testing data with the specified algorithm, the data preparation process first adds one label attribute to each class because the test is carried out using one of the classification methods in data mining, supervised learning.

The labels assigned to the dataset correspond to the data class on each type of pure tea tested. The four types of tea are numerically symbolized black tea with symbol 1, green tea with symbol 2, oolong tea with symbol 3, and white tea with symbol 4. In addition to the four types of pure tea used as data classes, the attributes of the data also need to be considered. The attributes in the dataset consist of four, namely CO2, CO, H2, and CH4. The four attributes are the content of the gas concentration present in the tea vapor.

The dataset displayed is raw data obtained from the data retrieval stage with an electronic nose device that has yet to be processed. The dataset that exists today is purely from taking electronic nose devices. In table 4, you can see the dataset that is ready to be tested.

Table 4. Four Attributes Dataset

No.	CO_2	CO	H_2	CH_4	Class
1.	302	428	25	67	1
2.	302	428	25	67	1
3.	302	428	25	67	1
4.	302	428	26	68	1
…	…	…	…	…	…
691.	276	377	25	97	2
692.	276	377	25	98	2
693.	276	377	25	97	2
694.	276	377	25	97	2
…	…	…	…	…	…
1379.	283	396	25	69	3
1380.	288	396	25	69	3
1381.	277	387	25	69	3
1382.	283	396	25	69	3
…	…	…	…	…	…
2753.	324	471	27	72	4
2754.	324	471	27	72	4
2755.	324	471	27	72	4
2756.	324	471	27	72	4

3.1. Dataset Visualization

The presentation of data or information is easier to understand through a graphical or visual display. Visual elements displayed to present data or information aim to provide convenience in understanding data trends, outliers, noise and missing values in the data. So that researchers are easier to see the condition of the data and decide on methods for handling data before a model enters them.

Figure 6 visualizes black tea data with its four attributes from the 1st to 689th data samples. The black tea data's visual display shows the data's unstable condition at the beginning of data retrieval by the sensor until the moment before the sensor stops taking data.

It is seen in the attributes for CO_2 and CO. The H_2 and CH_4 attributes appear stable from the beginning of data retrieval to the end of the sensor stopping to retrieve data.

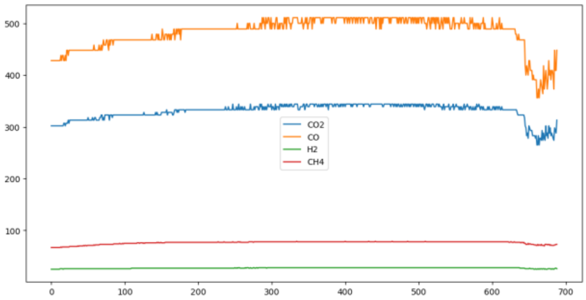


Figure 7. Black Tea Data Visualization

Figure 7 visualizes green tea data with its four attributes from the 1st to 689th data samples. The visual display of the green tea data shows a relatively stable data condition at the beginning of data retrieval by the sensor. Still, there is a decrease in CO_2 and CO attributes along the data interval but a drastic increase in values when the sensor stops taking data.

Data visualization does not show any outliers, noise or missing values in the dataset. This is seen in the attributes for CO_2 and CO. The H_2 and CH_4 attributes appear stable from the beginning of data retrieval to the end of the sensor stopping to retrieve data.

Figure 8 visualizes the oolong tea data with its four attributes from the 1st to the 689th data sample. The visual display of the oolong tea data shows a relatively stable data condition at the beginning of data retrieval by the sensor until before the sensor stops taking data. Data visualization does not show any outliers, noise or missing values in the dataset. This is seen in the attributes for CO_2 and CO. The H_2 and CH_4 attributes

DOI: https://doi.org/10.29207/resti.v7i2.4687

appear stable from the beginning of data retrieval to the end of the sensor stopping to retrieve data.

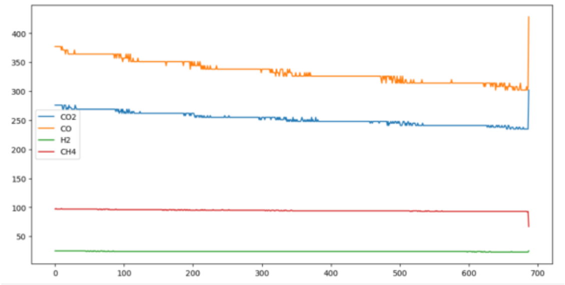


Figure 8. Green Tea Data Visualization

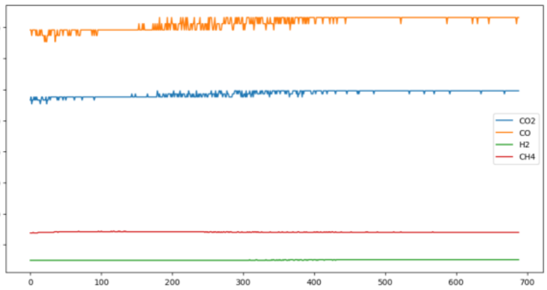


Figure 9. Oolong Tea Data Visualization

Figure 9 visualizes white tea data with its four attributes from the 1st to 689th data samples. The visual display of white tea data shows a relatively stable data condition at the beginning of data retrieval by the sensor until before the sensor stops taking data. This can be seen in the CO attribute. However, there was little noise at the beginning of the CO_2 attribute data collection. This is seen in the attributes for CO_2 and CO. The data visualization does not show any outliers or missing values in the dataset. The H_2 and CH4 attributes look very stable from the beginning of data retrieval to the end of the sensor stopping to retrieve data.

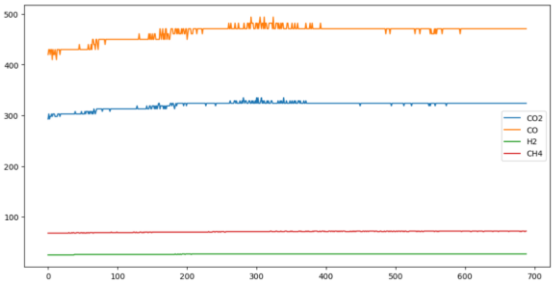


Figure 10. White Tea Data Visualization

Figure 10 visualizes white tea data with its four attributes from the 2067th to the 2756th data sample. The visual display of white tea data shows a pretty stable data condition at the beginning of data retrieval by the sensor until before the sensor stops taking data. This can be seen in the CO attribute. However, there is little noise at the beginning of the CO_2 attribute data collection. This is seen in the attributes for CO_2 and CO. The H_2 and CH_4 attributes appear stable from the birth of data retrieval to the end of the sensor stopping to retrieve data.

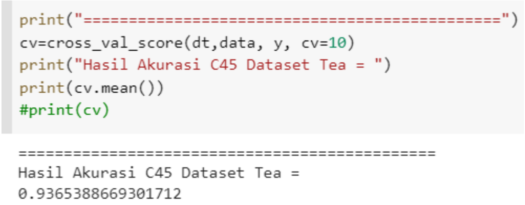
3.2. C4.5 Algorithm Calculation

The test was performed by testing a pure tea dataset tested with the C4.5 algorithm to measure the predictive accuracy of the refined tea dataset classification. Testing datasets with the C4.5 algorithm uses a 10fold cross-validation evaluation to generate accuracy from 10 more valid iterations. Researchers performed ten iterations that resulted in 10 test results. The results of the ten tests are then calculated on average to get the value of the prediction accuracy results. The results of the accuracy of 10fold cross-validation can be seen in Table 5.

Table 5. C4.5 Testing with 10fold Cross-Validation

k-fold	Accuracy Result
1	55,07%
2	96,73%
3	100%
4	100%
5	100%
6	99,63%
7	100%
8	100%
9	85.09%
10	85,09%

The average accuracy of testing the C4.5 algorithm with 10fold cross-validation resulted in a percentage of 93.65%. The results of the test accuracy of the C4.5 algorithm can be seen in Figure 11.


```
print("=============================================")
cv=cross_val_score(dt,data, y, cv=10)
print("Hasil Akurasi C45 Dataset Tea = ")
print(cv.mean())
#print(cv)
```

```
=============================================
Hasil Akurasi C45 Dataset Tea =
0.9365388669301712
```

Figure 11. C4.5 Algorithm Testing Accuracy Results

3.3. Chi-square & C4.5 Calculation

Manual testing of the dataset with the Chi-square method with a stage according to the formula using an α value of 5% or a significance level of 95% resulted in a Chi-square value on each attribute of the pure tea dataset. The amount of weight obtained from the calculation of the Chi-square value can be seen in Table 6.

Table 6. Chi-square Calculation Results

Data Attributes	Chi-Square Value
CO_2	27,65
CO	29,64
H_2	25,62
CH_4	25,63

DOI: https://doi.org/10.29207/resti.v7i2.4687

Chi-square calculations result in the ranking of Chi-square values from each attribute of a pure tea dataset. This section tests the dataset with two attributes with the highest value, namely the CO_2 and CO attributes. The selection of two attributes results from an experiment with the development of an increase in the accuracy of the C4.5 algorithm. Researchers conducted an assessment with the three highest Chi-square value attributes, namely CO2, CO, and H_2 attributes, with decreasing accuracy from the performance of the C4.5 algorithm. Table 7 shows the dataset's two attributes' test results with Chi-square.

Table 7. Dataset Two Attributes

No.	CO_2	CO	Class
1.	302	428	1
2.	302	428	1
3.	302	428	1
…	…	…	…
691.	276	377	2
692.	276	377	2
…	…	…	…
1380.	288	396	3
1381.	277	387	3
…	…	…	…
2755.	324	471	4
2756.	324	471	4

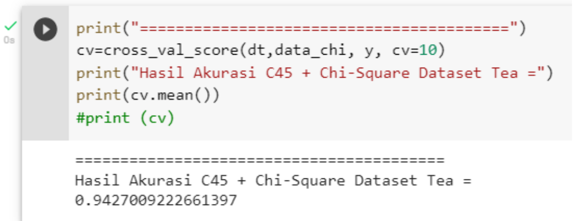
Testing the dataset with the Chi-square 2 attribute method and the C4.5 algorithm using 10fold cross-validation evaluation resulted in the accuracy of 10 iterations. The results of the accuracy of 10fold cross-validation can be seen in Table 8.

Table 8. Chi-C4.5 Testing with 10fold Cross-Validation

k-fold	Accuracy Result
1	68,11%
2	95,28%
3	94,56%
4	100%
5	99,27%
6	100%
7	100%
8	100%
9	100%
10	85,45%

The test results showed an increased accuracy from the previous test, namely testing the dataset with the C4.5 algorithm complete with 4 dataset attributes compared to testing the dataset using the C4.5 algorithm accompanied by Chi-square 2 attributes. CO_2 and CO attributes became two selected attributes tested with the C4.5 algorithm.

The average accuracy of testing the C4.5 algorithm with 10fold cross-validation resulted in a percentage of 94.27%. The results of the Chi-square and C4.5 test accuracy can be seen in Figure 12.



```
print("========================================")
cv=cross_val_score(dt,data_chi, y, cv=10)
print("Hasil Akurasi C45 + Chi-Square Dataset Tea =")
print(cv.mean())
#print (cv)
```

```
========================================
Hasil Akurasi C45 + Chi-Square Dataset Tea =
0.9427009222661397
```

Figure 12. Chi-Square and C4.5 Test Accuracy Results

3.4. Comparison of Calculation & Evaluation Results

The evaluation stage of the C4.5 algorithm describes the results of calculating the recall, precision, accuracy and F1-Score values of each class in the pure tea dataset before pre-processing. This stage is performed after the dataset is tested with the C4.5 algorithm. The confusion matrix values of the C4.5 algorithm test can be seen in Table 9.

Table 9. C4.5 Algorithm Testing Evaluation

		Predictive			
	Class	A	B	C	D
Actual	A	588	1	16	19
	B	13	687	0	36
	C	39	0	673	2
	D	49	0	0	632

The results of the calculation of recall, precision, accuracy and F1-score values from 2 classes in the tea dataset are presented in Table 10 for complete analysis and measuring the level of correctness on the performance of the Chi-square method and the C4.5 algorithm.

Table 10. C4.5 Algorithm Test Evaluation Results

Class	Recall	Precision	Accuracy	F1-Score
Black	0,942	0,85	0,95	0,896
Green	0,933	0,999	0,981	0,964
Oolong	0,942	0,977	0,98	0,96
White	0,928	0,917	0,961	0,922
Mean	0,93625	0,93575	0,935	0,9355

The evaluation stage of the Chi-square method and the C4.5 algorithm describes the calculation of each class's recall, precision, accuracy and F1-Score values in the pure tea dataset after pre-processing with the technique. The confusion values of the Chi-square testing matrix and the C4.5 algorithm can be seen in Table 11.

The results of the calculation of recall, precision, accuracy and F1-score values from 2 classes in the tea dataset are presented in Table 11 for comprehensive analysis and measuring the level of correctness on the performance of the Chi-square method and the C4.5 algorithm. The results of the values generated from the evaluation of the Chi-square test and the C4.5 algorithm can be seen in Table 12.

DOI: https://doi.org/10.29207/resti.v7i2.4687

Table 11. Chi-C4.5 Algorithm Testing Evaluation

		Predictive			
	Class	A	B	C	D
Actual	A	632	34	6	61
	B	7	564	0	0
	C	30	0	683	0
	D	20	0	0	628

Table 12. Chi-C4.5 Algorithm Test Evaluation Results

Class	Recall	Precision	Accuracy	F1-Score
Black	0,942	0,85	0,95	0,896
Green	0,933	0,999	0,981	0,964
Oolong	0,942	0,977	0,98	0,96
White	0,928	0,917	0,961	0,922
Mean	0,93625	0,93575	0,935	0,9355

Evaluation of confusion matrix calculations between testing datasets using the C4.5 algorithm with the Chi-square and C4.5 methods can be compared to the values of recall, precision, accuracy and F1-score measurements. The comparative value of such measures is shown in Table 13.

Table 13. Comparison of C4.5 and Chi-C4.5 Measurements

Evaluation	C4.5	Chi-C4.5
Recall	0,93625	0,940793
Precision	0,93575	0,944253
Accuracy	0,935	0,9427
F1-score	0,9355	0,941889

The calculation results of the two experimental scenarios showed the difference in the test results performed in the study between the testing of the C4.5 algorithm and the Chi-square C4.5 method on pure tea datasets. Using the number of attributes in an algorithmic test of a dataset affects the accuracy of classification predictions in the research conducted. Using two data attributes proves an increase in accuracy results compared to using four data attributes.

Graphic visualization of the comparison of the measurement of the values of the evaluation results of the C4.5 algorithm and the Chi-square C4.5 algorithm is shown in Figure 13.

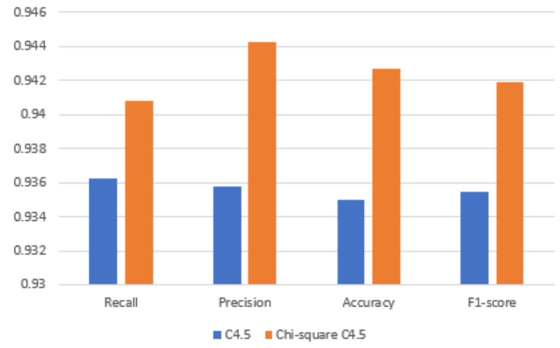


Figure 13. Graph Comparison of Testing C4.5 and Chi-C4.5

4. Conclusion

This research succeeded in developing an electronic nose device to retrieve pure tea datasets from dissolved tea vapors resulting in four data attributes and 2756 data records as objects to be tested with the C4.5 algorithm. The contribution of this study is to prove that the development of electronic nose devices for tea vapor can produce stable datasets to produce a high level of accuracy in testing datasets into classification methods.

In this study, testing the dataset with the C4.5 algorithm resulted in a classification prediction accuracy of 93.51%. Increasing accuracy in testing datasets with the Chi-square method improved the accuracy of the C4.5 algorithm by selecting two attributes and using two features with the highest Chi-square values, namely CO2 and CO features, resulting in an accuracy of 94.27%. The two data attributes with the highest Chi-square values selected in the dataset test resulted in higher accuracy than testing a dataset with four features. Meanwhile, the use of three data attributes with the highest Chi-square values results in a decrease in the accuracy rate in the classification of pure tea.

Based on the test results in this study in the classification of pure tea datasets with the C4.5 algorithm and the Chi-square method, it has reached a high level of accuracy. But there are still weaknesses that need to be redeveloped. The fault that needs to be redeveloped is that the dataset generated by the electronic nose device only consists of four data attributes from three sensors installed on the electronic nose device.

The conclusion drawn based on the tests carried out in this study is that weaknesses still need to be developed in selecting methods to help improve the accuracy of the C4.5 algorithm. Researchers suggested adding a gas sensor to the electronic nose to obtain more dataset attributes from the research conducted. In a later study, you can apply other methods or select areas other than attribute selection to help improve the accuracy of the C4.5 algorithm because Chi-square can only provide a 0.8% increase in the accuracy performance of the C4.5 algorithm with two selected attributes.

References

[1] D. Sita, Kralawi; Rohdinan, *Radar Opini dan Analisis Perkebunan*, 2nd ed. Bandung: dePlantation, 2021.

[2] B. P. Statistik, *Statistik Teh Indonesia*, 1st ed. Jakarta: Badan Pusat Statistik Republik Indonesia, 2018.

[3] Ditjenbun, *Buku Outlook Komoditas Perkebunan Teh*. Jakarta: Pusdatin Kementerian Pertanian, 2019.

[4] 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.

[5] S. Wakhid, R. Sarno, S. I. Sabilla, and D. B. Maghfira, “Detection and classification of indonesian civet and non-

civet coffee based on statistical analysis comparison using E-Nose," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 4, pp. 56–65, 2020, doi: 10.22266/IJIES2020.0831.06.

[6] A. Nugroho and Y. Religia, "Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 3, pp. 504–510, 2021, doi: 10.29207/resti.v5i3.3067.

[7] C. K. Lo, H. C. Chen, P. Y. Lee, M. C. Ku, L. Ogiela, and C. H. Chuang, "Smart dynamic resource allocation model for patient-driven mobile medical information system using C4.5 algorithm," *J. Electron. Sci. Technol.*, vol. 17, no. 3, pp. 231–241, 2019, doi: 10.11989/JEST.1674-862X.71018117.

[8] B. F. Tanyu, A. Abbaspour, Y. Alimohammadlou, and G. Tecuci, "Landslide susceptibility analyses using Random Forest, C4.5, and C5.0 with balanced and unbalanced datasets," *Catena*, vol. 203, Aug. 2021, doi: 10.1016/j.catena.2021.105355.

[9] X. Meng, P. Zhang, Y. Xu, and H. Xie, "Construction of decision tree based on C4.5 algorithm for online voltage stability assessment," *Int. J. Electr. Power Energy Syst.*, vol. 118, Jun. 2020, doi: 10.1016/j.ijepes.2019.105793.

[10] P. Gite, K. Chouhan, K. Murali Krishna, C. Kumar Nayak, M. Soni, and A. Shrivastava, "ML Based Intrusion Detection Scheme for various types of attacks in a WSN using C4.5 and CART classifiers," *Mater. Today Proc.*, Jul. 2021, doi: 10.1016/j.matpr.2021.07.378.

[11] S. Sundaramurthy and P. Jayavel, "A hybrid Grey Wolf Optimization and Particle Swarm Optimization with C4.5 approach for prediction of Rheumatoid Arthritis," *Appl. Soft Comput. J.*, vol. 94, p. 106500, 2020, doi: 10.1016/j.asoc.2020.106500.

[12] D. Marelli and M. Fu, "Asymptotic properties of statistical estimators using multivariate Chi-squared measurements," *Digit. Signal Process. A Rev. J.*, vol. 103, p. 102754, 2020, doi: 10.1016/j.dsp.2020.102754.

[13] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 2, pp. 225–231, 2020, doi: 10.1016/j.jksuci.2018.05.010.

[14] N. Peker and C. Kubat, "Application of Chi-square discretization algorithms to ensemble classification methods," *Expert Syst. Appl.*, vol. 185, no. July, p. 115540, 2021, doi: 10.1016/j.eswa.2021.115540.

[15] H. Chen, S. Fu, and H. Wang, "Optical coherence tomographic image denoising based on Chi-square similarity and fuzzy logic," *Opt. Laser Technol.*, vol. 143, no. July 2020, p. 107298, 2021, doi: 10.1016/j.optlastec.2021.107298.

[16] 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, p. 107301, 2021, doi: 10.1016/j.asoc.2021.107301.

[17] D. B. Magfira and R. Sarno, "Classification of Arabica and Robusta coffee using electronic nose," *2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018-Janua, pp. 645–650, 2018, doi: 10.1109/ICOIACT.2018.8350725.

[18] D. R. Wijaya, R. Sarno, and E. Zulaika, "Electronic nose dataset for beef quality monitoring in uncontrolled ambient conditions," *Data Br.*, vol. 21, pp. 2414–2420, 2018, doi: 10.1016/j.dib.2018.11.091.

[19] W. Harsono, R. Sarno, and S. I. Sabilla, "Recognition of original arabica civet coffee based on odor using electronic nose and machine learning," *Proc. - 2020 Int. Semin. Appl. Technol. Inf. Commun. IT Challenges Sustain. Scalability, Secur. Age Digit. Disruption, iSemantic 2020*, pp. 333–339, 2020, doi: 10.1109/iSemantic50169.2020.9234234.

[20] A. I. F. Al Isyrofie *et al.*, "Odor clustering using a gas sensor array system of chicken meat based on temperature variations and storage time," *Sens. Bio-Sensing Res.*, vol. 37, no. July, p. 100508, 2022, doi: 10.1016/j.sbsr.2022.100508.

[21] L. Ji, P. Liu, and S. Robert, "Tail asymptotic behavior of the supremum of a class of chi-square processes," *Stat. Probab. Lett.*, vol. 154, p. 108551, 2019, doi: 10.1016/j.spl.2019.07.001.