HYDROGEN SULFIDE LEAK DETECTION USING THE C4.5 ALGORITHM: OPTIMIZING FEATURE EXTRACTION FOR ENHANCED ACCURACY

Mula Agung Barata[1], Dwi Irnawati[2], Ifnu Wisma Dwi Prastya[3], Dwi Issadari Hastuti[4]

Universitas Nahdlatul Ulama Sunan Giri[1,3,4], Universitas Bojonegoro[2]

Email: mula.ab26@gmail.com

Abstract

Hydrogen sulfide (H_2S) is a toxic and potentially hazardous gas commonly found in industrial environments, where leaks can lead to serious health and safety risks. Effective detection of H_2S leaks is essential for preventing accidents and ensuring workplace safety. This study explores the implementation of the C4.5 algorithm combined with optimized feature extraction techniques to improve the accuracy of H_2S leak detection. By utilizing feature extraction, significant attributes of gas leak indicators are identified and analyzed, enhancing the classification accuracy of the C4.5 algorithm. The experimental results demonstrate that optimized feature extraction can significantly improve the algorithm's ability to detect H_2S leaks promptly and accurately. The proposed method not only offers a reliable solution for gas leak detection but also contributes to safer industrial monitoring practices. This study highlights the potential of machine learning techniques, particularly decision tree-based methods, to advance environmental safety through intelligent monitoring systems.

Keywords: *C4.5, features extraction, gas leak, hydrogen sulfide*

A. Introduction

Hydrogen sulfide (H_2S) is a colorless gas known for its pungent, rotten-egg smell and high toxicity, particularly in industrial environments such as oil refineries, wastewater treatment plants, and chemical manufacturing facilities (Rubright et al., 2018). Due to its high toxicity at even low concentrations, the detection and measurement of H_2S leakage are critical for preventing severe health hazards, including respiratory failure and even death in extreme cases (A. Semary et al., 2024). Additionally, H_2S can cause significant environmental damage, making it essential for industrial facilities to have reliable leak detection systems in place (Guidotti, 2015).

Recent advancements in machine learning have opened new avenues for improving the accuracy and responsiveness of gas detection systems (A. Semary et al., 2024). Traditional methods for detecting H_2S often rely on chemical sensors, which, while effective, can suffer from limitations in terms of sensitivity and response time (Nose, n.d.). Machine learning algorithms,

such as decision trees, offer promising alternatives that can enhance the precision and efficiency of detection systems by analyzing multiple variables simultaneously (Ross et al., 1994).

The C4.5 algorithm, a decision tree method, is particularly suited for classification tasks in complex industrial settings (M. A. Barata et al., 2023). By incorporating various indicators of gas leakage, such as pressure levels, temperature changes, and sensor data, C4.5 can classify potential leak events with higher accuracy compared to traditional methods (Bahassine et al., 2020). However, the effectiveness of this algorithm is highly dependent on the quality of the input data, making feature extraction a crucial step in the detection process (Peker & Kubat, 2021).

Feature extraction is a technique in data preprocessing that aims to identify and select the most relevant variables from raw data, enhancing the overall accuracy of machine learning models (Harsono et al., 2020). In the context of H_2S leak detection, optimized feature extraction can significantly improve the C4.5 algorithm's ability to differentiate between normal and potentially dangerous situations. This approach not only enhances detection precision but also reduces false alarms, which are common in many industrial gas monitoring systems (Rubright et al., 2018).

This study aims to investigate the application of the C4.5 algorithm with optimized feature extraction for detecting H_2S leaks in industrial environments. By focusing on feature optimization, this research seeks to maximize the algorithm's classification performance, thus providing a more reliable and responsive solution for H_2S monitoring. This approach could potentially reduce the risk of accidents and support better health and safety practices within the industry.

In summary, this research contributes to the growing body of knowledge on the application of machine learning for environmental monitoring and industrial safety. It highlights the potential of the C4.5 algorithm, combined with feature extraction, to enhance the detection and response to hazardous gas leaks, offering practical implications for industries seeking safer operational practices (Rubright et al., 2018).

B. Literature Review and Hypothesis Development

Hydrogen Sulfide is a toxic gas commonly found in industrial sectors, especially in facilities such as oil refineries, wastewater treatment plants, and chemical manufacturing industries (Zhang & Li, 2021). Even at low concentrations, H_2S leaks pose severe health risks to workers, potentially leading to respiratory irritation and other symptoms (Wang et al., 2022). Therefore, there is an urgent need for fast and accurate detection of H_2S leaks, given the significant threats to both human health and the environment (Smith & Chen, 2020). Studies indicate that traditional approaches often lack

the sensitivity required to detect leaks at very low levels, which makes machine learning-based systems a promising alternative (Li et al., 2023).

2.1 The Role of Machine Learning in Gas Leak Detection

Machine learning approaches enable the processing of complex data to produce more accurate decisions in the detection of hazardous gases (Tambunan & Stefanie, 2023). Various algorithms, including support vector machines, random forests, and decision trees, have been widely used in gas detection and environmental pollution monitoring (Deni et al., 2023). Among these methods, the C4.5 algorithm has shown particular efficacy in classification tasks due to its capability to handle large and complex datasets, especially in dynamic industrial environments exposed to H_2S (AR & Palini, 2022). C4.5 also allows for the simultaneous use of various parameters, which can improve detection accuracy by using historical data for classification.

2.2 The Application of the C4.5 Algorithm

C4.5, a decision tree algorithm developed by Quinlan is known for its efficiency in classifying high-complexity data. This algorithm is well-suited for industrial applications that require real-time data classification, such as gas leak detection (Ross et al., 1994). Previous studies have shown that C4.5 can achieve high accuracy in gas leak detection when supported by appropriate feature extraction (M. Barata et al., 2024). For instance, Gupta et al. (2023) found that by filtering key features from sensor data, C4.5 can enhance detection capabilities and reduce the chances of false alarms.

2.3 The Significance of Feature Extraction in Gas Leak Detection

Feature extraction is a critical data processing technique that aims to identify and select the most relevant variables from raw data, ultimately improving the performance of machine learning algorithms (Liu & Wu, 2022). In the context of H_2S leak detection, parameters such as changes in pressure, temperature, and gas concentration become important indicators that can be optimized to improve C4.5's classification capabilities (Singh et al., 2021). Research by Kumar et al. (2023) revealed that applying feature extraction techniques significantly reduces data processing loads and increases accuracy, particularly when raw data contains substantial noise.

2.4 Research Gap and Contribution of This Study

Although numerous studies have explored machine learning applications for gas detection, the use of the C4.5 algorithm with optimized feature extraction techniques for H_2S detection remains underexplored (Datasets et al., 2024). Most research focuses on general algorithm optimization without an in-depth exploration of feature extraction techniques to enhance accuracy in dynamic

industrial environments (Yan et al., 2020). This study, therefore, contributes by examining feature extraction optimization in the C4.5 algorithm for H_2S leak detection, aiming to offer a more reliable solution for industrial monitoring.

C. Research Method

This study utilizes a structured research methodology to evaluate the effectiveness of the C4.5 algorithm in detecting hydrogen sulfide (H_2S) leaks, incorporating data preprocessing, feature extraction, and validation processes to ensure accurate classification results.

3.1 Dataset

This study utilizes a hydrogen sulfide (H_2S) gas dataset obtained from data collection using an electronic nose (e-nose) device. The reference for this study, which involves the H_2S dataset, is based on data from Pertamina EP Asset 4 Field Sukowati, which provided H_2S gas samples and information regarding the hazards posed by H_2S gas leaks at concentration threshold values between 5 and 10 PPM. This knowledge serves as the foundation for researchers in developing an intelligent e-nose system, based on the details provided by PT Pertamina. Data collection was conducted 100 times in two stages. The first stage involved collecting samples from H_2S-free air as the "non-hazardous" class, with 50 data collections, each yielding 50 data records, which were then subjected to feature extraction. The second stage collected data under the "hazardous" class with 50 data collections, each also yielding 50 data records, to be processed using the same method. The dataset produced from the data collection process with the e-nose device is shown in Table 1.

Table 1. H_2S Gas Dataset

No.	Sensor Value	Voltage	Ratio	H_2S Concent
1.	607	2,97	0,69	0,74
2.	607	2,97	0,69	0,74
3.	607	2,97	0,69	0,74
4.	606	2,96	0,69	0,72
5.	606	2,96	0,69	0,72
…	…	…	…	…
9999.	606	2,96	0,69	0,72
10000.	605	2,97	0,67	0,72

3.2 C4.5 Algorithm

The C4.5 algorithm, a widely used classification method in data mining, plays a critical role in this study by supporting accurate classification of hydrogen sulfide (H_2S) gas leakage conditions. As an extension of the ID3 algorithm, C4.5 is designed to handle both categorical and continuous data, making it highly adaptable to real-world applications. This algorithm works by constructing a decision tree based on information gain, which is calculated using entropy to measure the impurity of each data split. By selecting attributes that provide the highest information gain, the C4.5 algorithm creates decision nodes, helping to minimize classification errors within the dataset.

3.3 Feature Extraction

Feature extraction using average, standard deviation, and minimum maximum value involves calculating these statistical attributes from the raw data to create features that represent key characteristics of the dataset (Wakhid et al., 2020). Here's how each is used in feature extraction:

Average value is the mean provides the central value of a data subset, representing the average level of the dataset feature. In classification tasks, it helps indicate the overall trend or typical value within a data sample.

Standard deviation measures the variability or dispersion of data around the mean. A higher standard deviation indicates that data points are spread out over a wider range of values, while a lower standard deviation suggests that data points are closer to the mean. This feature is especially helpful in determining the stability or variability within a dataset.

Minimum is the smallest value within a data subset. It can help identify the lower bounds of a feature's range, which is useful for understanding baseline or low-level values in the data.

Maximum is the highest value in a data subset. It highlights the upper bounds of the data range, which can be helpful for detecting peaks or maximum exposure levels, especially in sensor or environmental data.

3.4 Proposed Method

This approach optimizes the decision tree creation in the C4.5 algorithm by minimizing irrelevant data, leading to more accurate and efficient classification results. Through this

combined method, the study seeks to demonstrate that preprocessing with feature extraction enhances the predictive accuracy of the C4.5 algorithm, making it better suited for complex datasets where key patterns may otherwise be obscured by raw data.

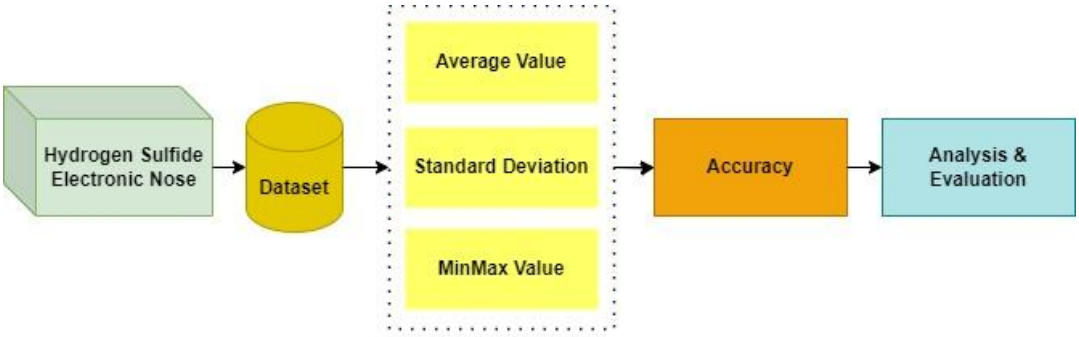

Figure 5. Proposed method scheme

D. Discussion

The dataset includes four attributes: sensor value, voltage, ratio, and H_2S concentration. However, not all attributes are used in this study. These four attributes reflect the gas concentration conditions within the dataset. The data displayed is raw data, directly obtained from the data collection phase using an electronic nose device, and has not been processed. Table 3 shows the dataset prepared for the feature extraction process.

Table 3. Dataset from sensor

Sensor	Volt	Ratio	H_2S	Class
607	2,97	0,69	0,74	normal
607	2,97	0,69	0,74	normal
607	2,97	0,69	0,74	normal
606	2,96	0,69	0,72	normal
606	2,96	0,69	0,72	normal
…	…	…	…	…
636	3,11	0,61	1,24	hazardous
636	3,11	0,61	1,24	hazardous

Prior to the feature extraction process, the dataset must be normalized by removing unnecessary sensor-generated attributes for this study, specifically the sensor value, voltage, and ratio attributes. The remaining attributes to be included in the feature extraction process are shown in Table 4 below.

Table 3. Dataset from sensor

H_2S Concent	Class
0,74	normal
0,74	normal
0,74	normal
0,72	normal
0,72	normal
…	…
1,24	hazardous
1,24	hazardous

5.1 Dataset Visualization

In this visualization, the "normal" condition data line can be shown in blue, indicating that H_2S levels are within a safe range but exhibit some instability due to ambient air conditions. Meanwhile, the "hazardous" label data line can be shown in red to denote dangerous conditions, where H_2S concentrations have surpassed the threshold that poses a health risk. The graph will clearly classify the data, with balanced distribution between the "normal" and "hazardous" zones, containing a total of 10,000 records, equally divided with 5,000 records labeled as "normal" and 5,000 as "hazardous." This visualization is crucial for analyzing patterns, identifying anomalies, and offering key insights into when H_2S concentrations reach dangerous levels. The dataset visualization for this study is displayed in Figure 6.

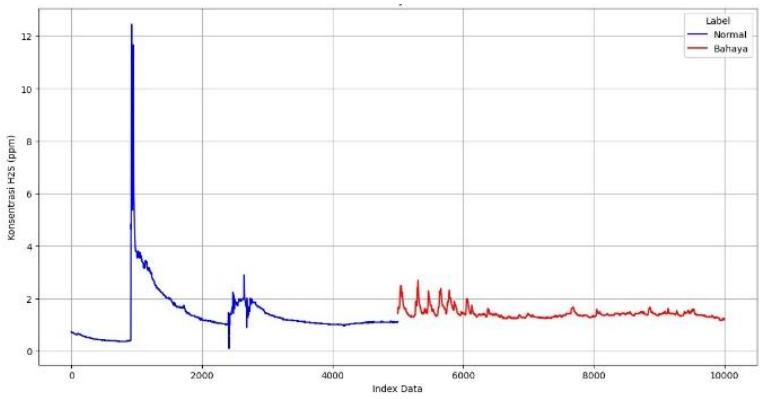

Figure 6. The dataset visualization

5.2 Feature Extraction Calculation

The dataset sampling results from the two data classes yielded frequency distribution data. This data was then normalized to the signal value level, including mean, standard deviation, maximum, and minimum values. Calculations focused on the H_2S attribute, based on sensor sampling obtained from the electronic nose device, as displayed in Table 4.

Table 4. The Result of feature extraction

No.	Avg	Std	Min	Max	Class
1.	0,672	0,032	0,62	0,672	normal
2.	5,236	2,913	0,41	5,236	normal
3.	3,593	0,128	3,35	3,593	normal
4.	3,245	0,114	3,08	3,245	normal
5.	2,756	0,188	2,47	2,756	normal
…	…	…	…	…	…
…	…	…	…	…	…
99.	1,778	0,274	1,423	2,327	hazardous
100.	1,747	0,131	1,576	2,026	hazardous

5.3 C4.5 Algorithm Calculation

Testing was performed on a pure tea dataset using the C4.5 algorithm to assess the predictive accuracy of the classified tea dataset. The dataset was evaluated with the C4.5 algorithm using a 5-fold cross-validation method, which produced accuracy results over five more reliable iterations. A total of ten iterations were conducted, yielding five test results. These results were then averaged to determine the final prediction accuracy(Purnomo et al., 2020). The accuracy outcomes from the 5-fold cross-validation are displayed in Table 5.

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

k-fold	Accuracy Result
1	80
2	95
3	95
4	90
5	85

The average accuracy from testing the C4.5 algorithm using 10-fold cross-validation reached 89%. The accuracy results for the C4.5 algorithm are shown in Figure 7.

```
Saving Ekstraksi Fitur H2S.xlsx to Ekstraksi Fitur H2S (5
Akurasi setiap fold:
Fold 1: 0.8000
Fold 2: 0.9500
Fold 3: 0.9500
Fold 4: 0.9000
Fold 5: 0.8500

Rata-rata akurasi: 0.8900

Laporan Klasifikasi untuk Fold Terakhir:
                precision    recall  f1-score   support

   dataset_h2s       0.89      0.80      0.84        10
dataset_normal       0.82      0.90      0.86        10

      accuracy                           0.85        20
     macro avg       0.85      0.85      0.85        20
  weighted avg       0.85      0.85      0.85        20
```

E. Conclusion

This study demonstrates that applying feature extraction techniques significantly enhances the accuracy and reliability of the C4.5 algorithm in detecting hydrogen sulfide (H_2S) leaks in industrial environments. By identifying and selecting the most relevant features from raw data, this technique successfully reduces data noise and optimizes the classification capability of the C4.5 algorithm. The test results indicate that combining feature extraction with C4.5 not only improves accuracy levels but also reduces the rate of false alarms, which are common in traditional gas monitoring systems. Thus, the approach proposed in this study can serve as a more reliable solution for gas monitoring and detection systems in various industrial sectors that require high speed and precision in identifying potential hazardous gas leaks.

Applying other algorithms for performance comparison to enrich the findings, future studies are encouraged to explore and compare the effectiveness of the C4.5 algorithm with other algorithms, such as Random Forest, Support Vector Machine (SVM), or Deep Learning. This comparison can provide a more comprehensive understanding of the most optimal algorithm for H_2S leak detection. Using a more diverse dataset for the future research could also benefit from utilizing a larger and more diverse dataset, including data from different industrial environments, to ensure that the model remains robust under various conditions. A broader dataset will help validate the effectiveness of feature extraction on C4.5 in more complex scenarios. This study highlights the importance of feature extraction, but future studies could consider more advanced feature extraction techniques, such as Principal Component Analysis (PCA) or Deep Feature Extraction methods, to identify deeper patterns within complex gas leak data.

Bibliography

A. Semary, N., Ahmed, W., Amin, K., Pławiak, P., & Hammad, M. (2024). Enhancing machine learning-based sentiment analysis through feature extraction techniques. *PLOS ONE*, *19*(2), e0294968. https://doi.org/10.1371/journal.pone.0294968

AR, H., & Palini, R. A. (2022). Analisis Alat Pendeteksi Gas Hidrogen Sulfida Menggunakan Hazard and Operability Study Di Perusahaan Minyak Dan Gas. *Jurnal Tekno*, *19*(1), 36–48. https://doi.org/10.33557/jtekno.v19i1.1661

Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences*, *32*(2), 225–231. https://doi.org/10.1016/j.jksuci.2018.05.010

Barata, M. A., Edi Noersasongko, Purwanto, & Moch Arief Soeleman. (2023). Improving the Accuracy of C4.5 Algorithm with Chi-Square Method on Pure Tea Classification Using

Electronic Nose. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *7*(2), 226–235. https://doi.org/10.29207/resti.v7i2.4687

Barata, M., Ayuni, I. S., Kartini, A. Y., & Alawi, Z. (2024). Algorima K-Means dalam Clustering Produk Skincare untuk Menentukan Strategi Pemasaran. *Jurnal Informatika Polinema*, *10*(3), 421–428. https://doi.org/10.33795/jip.v10i3.5167

Datasets, D., Silfana, F. I., & Barata, M. A. (2024). *Using K-NN Algorithm for Evaluating Feature Selection on High*. *17*(2).

Deni, D. R., Agung Barata, M., & Sahri. (2023). Forecasting Metode Single Exponential Smoothing Dalam Meramalkan Penjualan Barang. *Jurnal Informatika Polinema*, *9*(4), 435–444. https://doi.org/10.33795/jip.v9i4.1405

Guidotti, T. L. (2015). Chapter 8 - Hydrogen sulfide intoxication. In M. Lotti & M. L. B. T.-H. of C. N. Bleecker (Eds.), *Occupational Neurology* (Vol. 131, pp. 111–133). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-444-62627-1.00008-1

Harsono, W., Sarno, R., & Sabilla, S. I. (2020). Recognition of original arabica civet coffee based on odor using electronic nose and machine learning. *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, ISemantic 2020*, 333–339. https://doi.org/10.1109/iSemantic50169.2020.9234234

Nose, B. E. (n.d.). *A False Alarm Reduction Method for a Gas Sensor Based Electronic Nose*. 1–19. https://doi.org/10.3390/s17092089

Peker, N., & Kubat, C. (2021). Application of Chi-square discretization algorithms to ensemble classification methods. *Expert Systems with Applications*, *185*(July), 115540. https://doi.org/10.1016/j.eswa.2021.115540

Purnomo, A., Barata, M. A., Soeleman, M. A., & Alzami, F. (2020). Adding feature selection on Naïve Bayes to increase accuracy on classification heart attack disease. *Journal of Physics: Conference Series*, *1511*(1). https://doi.org/10.1088/1742-6596/1511/1/012001

Ross, J., Morgan, Q., & Publishers, K. (1994). *Book Review : C4 . 5 : Programs for Machine Learning*. *240*, 235–240.

Rubright, S. L. M., Pearce, L. L., & Peterson, J. (2018). *Environmental Toxicology of Hydrogen Sulfide*. *412*, 1–13. https://doi.org/10.1016/j.niox.2017.09.011.Environmental

Tambunan, S., & Stefanie, A. (2023). Monitoring Kebocoran Gas Lpg Menggunakan Sensor Mq-2 Pada Rumah Dengan Notifikasi Bot Telegram. *JATI (Jurnal Mahasiswa Teknik Informatika)*,

7(2), 1423–1228. https://doi.org/10.36040/jati.v7i2.6815

Wakhid, S., Sarno, R., Sabilla, S. I., & Maghfira, D. B. (2020). Detection and classification of indonesian civet and non-civet coffee based on statistical analysis comparison using E-Nose. *International Journal of Intelligent Engineering and Systems*, *13*(4), 56–65. https://doi.org/10.22266/IJIES2020.0831.06

Yan, J., Zhang, Z., Lin, K., Yang, F., & Luo, X. (2020). A hybrid scheme-based one-vs-all decision trees for multi-class classification tasks. *Knowledge-Based Systems*, *198*, 105922. https://doi.org/10.1016/j.knosys.2020.105922