Comparison of k-NN and Naive Bayes Algorithms for Classifying Mackerel Tuna Freshness For Real-Time Classification Using Gas Sensors

1st Muhammad Rafi Mahfuz Setyagraha
School of Electrical Engineering
Telkom University
Bandung, Indonesia
rafimahfuz@student.telkomuniversity.ac.id

2nd Ledya Novamizanti
School of Electrical Engineering
Telkom University
Bandung, Indonesia
ledyaldn@telkomuniversity.ac.id

3rd Dedy Rahman Wijaya
School of Applied Science
Telkom University
Bandung, Indonesia
dedyrw@telkomuniversity.ac.id

***Abstract*—The large production and consumption of mackerel tuna in Indonesia reflect its importance as a local staple and a valuable export product contributing to the nation's economy. Mackerel tuna is prized for its nutritional content and affordability, making it a crucial part of the diet for many Indonesians. Ensuring the freshness and quality of this high-demand product is essential. This study introduces a machine-learning approach to detect fish freshness by analyzing gases emitted during spoilage, utilizing MQ-2, MQ-9, and MQ-135 gas sensors. The data were processed using the k-Nearest Neighbors (k-NN) and Naive Bayes algorithms, both achieving accuracy rates near 100%. These findings highlight the system's potential to enhance quality control in Indonesia's fishery industry by offering an efficient and reliable method for assessing fish freshness.**

***Keywords*—classification, machine learning, tuna, gas sensor**

I. INTRODUCTION

The fisheries sector is a significant component of Indonesia's economy, contributing substantially to national and agricultural GDP. In 2022, the industry accounted for approximately 555 billion IDR, 2.6% of the national and 21% of the agricultural GDP [1]. It employs approximately six million individuals, underscoring its role in livelihood and economic stability [2]. Additionally, per capita fish consumption in Indonesia reached 35.26 kilograms per year [3], highlighting the pivotal role of fish in the country's food security. Indonesia's fishery sector demonstrated its global significance in 2023, with a projected export value of $7.66 billion based on government targets [4]. However, the export figures may not have fully met these aspirations [5].

In 2021, Indonesia's fisheries sector contributed approximately 2.7% to the nation's total exports, valued at USD 5.15 billion [1]. The industry is crucial for the Indonesian economy, and maintaining high-quality fishery products is essential to avoid financial losses. Quality control for fishery products is predominantly performed manually, which is time-consuming and labor-intensive [6]. This manual process is time-consuming and prone to a 5-10% risk of human error, potentially compromising product quality. Non-compliance with international quality standards can lead to significant economic losses for exporters, as evidenced by product rejections in major markets like the United States. These rejections can result in substantial financial penalties, with estimates suggesting a significant economic impact. Additionally, poor product quality can damage the reputation of producers and shrink demand in the long run from global markets [7]. Ensuring high quality and safety standards is economically beneficial, as it minimizes losses from spoilage and trade disruptions. Therefore, implementing advanced and accurate technologies for assessing fishery product quality is vital.

The fisheries sector is vital to Indonesia's economy and is crucial to global food security. Seafood is a significant source of protein for millions worldwide. Ensuring the freshness and quality of fish products is paramount, as non-compliance with international standards can lead to substantial economic losses for exporters and affect global supply chains.

An alternative method for assessing fishery product quality involves utilizing the gas these products release. The concentration of gas released from the spoiling process of fish can be determined using a gas sensor. Electronic noses (e-noses), consisting of sensors that detect odors, have been successfully used in numerous studies to identify the freshness of beef and pork [8]. This technology offers a promising solution for evaluating the quality of seafood products.

Based on the background outlined, this research aims to utilize a gas sensor array system to assess the quality of seafood, specifically mackerel tuna, by analyzing the gases released from these fish. The study proposes training an algorithm that can rapidly classify the freshness of mackerel tuna, with a target accuracy of at least 98%. The research utilized a dataset comprising 36,032 records. To achieve this, experiments were conducted using four different types of gas sensors. The e-nose system was integrated with machine learning algorithms, employing hyperparameter optimization and noise filtering techniques to enhance accuracy in evaluating seafood freshness. The k-Nearest Neighbors (k-NN) and Naive Bayes algorithms were selected for this study based on previous research demonstrating their effectiveness in classification tasks, particularly in sensor data scenarios. K-NN is known for its simplicity and robustness in handling multi-class problems.

At the same time, Naive Bayes offers a fast and lightweight alternative, making both suitable for real-time applications in detecting seafood freshness.

II. RELATED WORKS

Incorporating sensors and machine learning technology for fish quality detection has become a crucial area of research toward effective and trustworthy seafood quality assessment systems. This study owes to its unique approach and more detailed progress retrieval [9], concentrating on well-diversified sensors, machine learning algorithms, and fish freshness estimation. Concerning this aspect, another review [10] focuses on the latest trends in sensing configurations to determine fish shelf-life along with the merits, downsides, and synergies of different sensors – gas, colorimetric – coupled with machine learning constructs. Another research [11] provides a detailed and high-level application of the electronic nose supported with advanced machine learning assisted by hyperparameter tuning for seafood quality detection. This study established that the most accurate algorithm was the k-Nearest Neighbors (k-NN) model, which recorded a classification accuracy of 1.0 and the best-performing model in regression tasks, achieving an RMSE of 0.003 and an R^2 of 0.99 in the detection of microbes in seafood. This emphasizes the applicability level of the e-nose and k-NN in seafood quality evaluation.

While previous studies have explored electronic noses and machine learning for seafood quality detection, this research distinguishes itself by employing a comprehensive gas sensor array and optimizing algorithm performance through hyperparameter tuning and noise filtering techniques.

III. METHODOLOGY

A. *Dataset Acquisition*

The device used in this experiment, shown in Fig. 1, consists of a gas sensor array connected to an ESP Devkit-C S3 microcontroller. The gas sensor array, the combination of which is detailed in Table I, is embedded in a semi-airtight container that serves as the sample chamber. The sample used in the experiment was procured fresh on the morning of July 22nd and immediately placed in the sample chamber. An algorithm periodically captures the data from the electronic nose and stores it in a Pandas DataFrame. The sampling process was conducted for 10 hours, recording data as the sample transitioned from fresh to spoiled at room temperature (25°C). Data was captured once every second, resulting in over 36,000 records saved in CSV format.

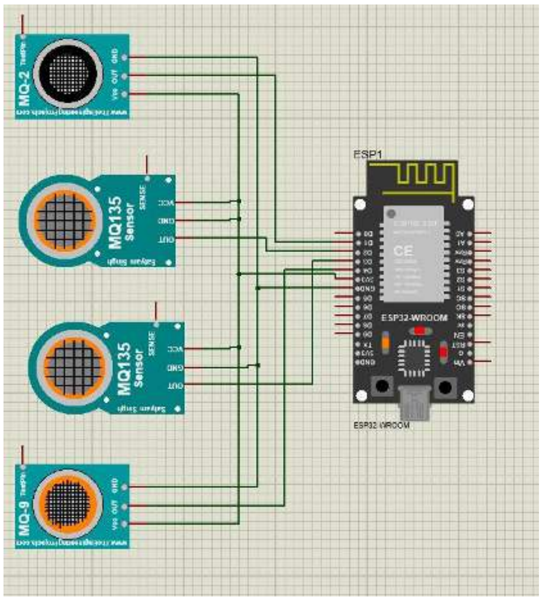

Fig. 1. E-nose device circuit diagram

In this study, an experimental setup was designed to evaluate the effectiveness of the k-Nearest Neighbors (k-NN) and Naive Bayes algorithms in classifying the freshness of mackerel tuna. The gas sensors used, as detailed in Table 1, include two MQ-135 sensors, one calibrated for Ammonia and the other for carbon dioxide, along with an MQ-2 sensor calibrated for Alcohol and an MQ-9 sensor calibrated for Methane. These sensors detect the compounds released during fish spoilage.

The following measures were taken to ensure the reliability and validity of the experimental results. To ensure the stability of the output data from the sensors, the sensors were pre-heated for more than 24 hours before use. In data preprocessing, raw and collected data were cleaned per the requirement for quality data. The data was split into 80% training and 20% testing data.

TABLE I. GAS SENSORS USED.

No.	Gas Sensor	Selectivity
1.	MQ-135	Ammonia
2.	MQ-2	Alcohol
3.	MQ-9	Methane
4.	MQ-135	Carbon Dioxide

The collected data were then labeled based on the time elapsed from the beginning of the sampling process, using a Python script. The labeling criteria were based on the time the samples remained at room temperature, categorizing them into three distinct freshness levels: The samples with an exposure time of fewer than 5 hours are labeled as – "Fresh"; the samples with an exposure time of between 5 and 7 hours are labeled as "Less Fresh"; and samples with an exposure time exceeding 7 hours are labeled as "Not Fresh/Reject.". The criteria used in the labeling process were provided by an expert in fish quality control who has worked in the field for over 19 years.

B. *Proposed Method*

This study compares the k-Nearest Neighbors (k-NN) and Naive Bayes algorithms to determine the most suitable model for predicting the freshness of mackerel tuna. The raw data obtained from the sensors required little data cleaning. The collected sensor data is divided into 80%

training and 20% test sets. The 80% training data is utilized for training the k-NN and Naive Bayes models, which also undergo hyperparameter tuning using the grid search method. For the k-NN algorithm, three fundamental parameters are optimized: the number of neighbors (k), the weights used in prediction, and the algorithm to find the nearest neighbors. The last 20% of the dataset is utilized to assess the accuracy of the models in classifying the freshness of mackerel tuna. The primary purpose of this approach is to find an efficient, accurate, and lightweight algorithm for fish freshness classification using gas sensors, which can be used as an effective and reliable quality control process.

Fig. 2 presents the workflow of the proposed method for the machine-learning aspect of this research. It outlines critical stages, including model training with hyperparameter optimization, performance testing, and evaluation. The process begins with selecting and preparing the dataset and then splitting the data into training and testing sets. During model training, hyperparameter optimization is employed to fine-tune the algorithms for optimal performance. Once trained, the model is tested on the validation set to assess its accuracy and generalization capability. Finally, the model's performance is evaluated, ensuring a thorough assessment of its effectiveness in classifying fish freshness.

The evaluation metrics used in the evaluation are accuracy, precision, recall, F1 score, and confusion matrix. Further, 5-fold cross-validation is used to estimate model accuracy by dividing the dataset into five subsets, in which only one subset is used for validation. In comparison, the other four subsets are used for training. The performance measure is averaged over all five splits, improving the reliability of the evaluation and reducing the effects of how the data is split. These steps are beneficial in providing a reliable comparison between the two models.

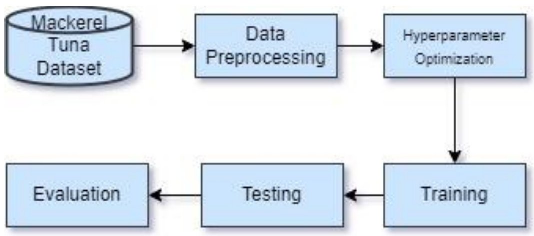

Fig. 2. Proposed method.

A method was also developed to evaluate both algorithms' average classification speed and memory usage. Both algorithms were used to classify 100 random data points from the test set, and the test results were plotted in a line graph. Lastly, the average classification time and memory usage were calculated.

IV. RESULTS AND ANALYSIS

The performance of the k-NN and Naive Bayes models was compared to analyze their ability to detect the freshness level of mackerel tuna using gas sensors. The performance metrics used in this analysis include 5-fold cross-validation, accuracy, precision, recall, F1 score, classification speed, and memory usage. Table II shows the parameters tested using the grid search method and the best parameters found.

TABLE II. GRID SEARCH PARAMETERS.

Algorithm	Parameters	Parameter Values	Best Parameters
k-NN	n_neighbors	5	10
		10	
		20	
		50	
		100	
		150	
	weights	uniform	uniform
		distance	
	algorithm	auto	auto
		ball_tree	
		kd_tree	
		brute	
		linear	
Naive Bayes	var_smoothing	1e -9	1e -9
		1e -8	
		1e -7	
		1e -6	
		1e -5	
		1e -4	
		1e -3	
		1e -2	
		1e -1	
		1	
		10	
		100	

The k-NN model's exceptional performance can be seen in its confusion matrix, shown in Fig 3. Only one misclassification out of 7,207 samples resulted in an accuracy rate of almost 100%. This was further enforced by the average cross-validation score of 0.99916, which shows the model's strong generalization ability. The precision, recall, and F1-score performance metrics for the k-NN model were also remarkably high, which attests to its strong classification capability.

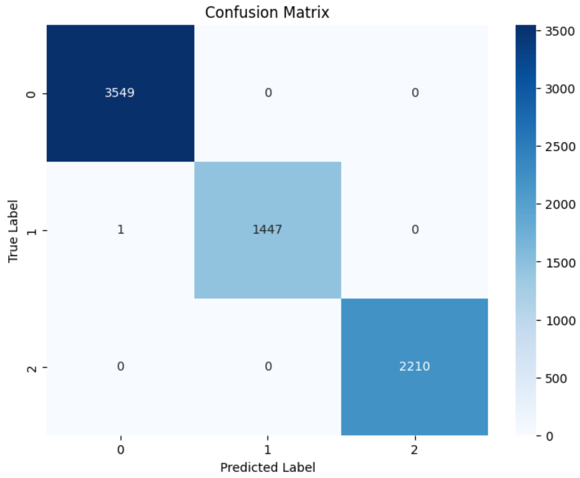

Fig. 3. k-NN model confusion matrix

In comparison, the Naive Bayes model also performed reasonably well. However, with lower accuracy, it misclassified fifty-four samples out of 7,207, as shown in its confusion matrix in Fig 4, resulting in an accuracy percentage of 99.2%. The average cross-validation score was 0.982, indicating solid generalization ability but lower than k-NN. The Precision, recall, and F1-score performance metrics for the Naive Bayes model also showed good reliability in classification tasks.

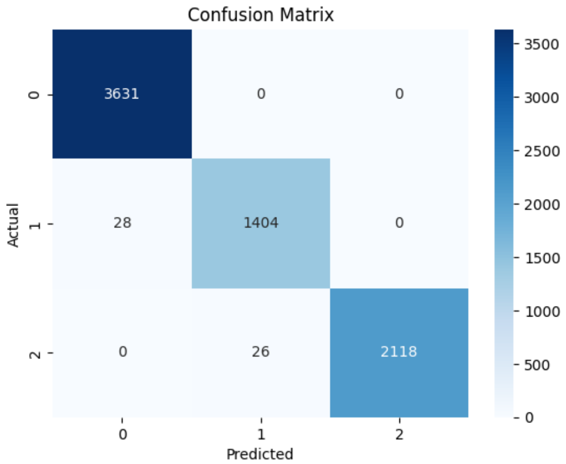

Fig. 4. Naive Bayes model confusion matrix

To evaluate both algorithms' classification time and memory usage, a testing method was developed that involves classifying 100 randomly selected data points from the test set. The results, which are plotted in Fig. 5 and summarized in Table III, indicate that while the Naive Bayes algorithm exhibits slightly lower accuracy and cross-validation scores compared to the k-NN algorithm, it significantly outperforms k-NN in terms of average classification time and memory usage. This suggests that despite its marginally lower accuracy, Naive Bayes is more efficient for applications with critical speed and memory efficiency. In contrast, k-NN may be better suited for scenarios prioritizing classification accuracy.

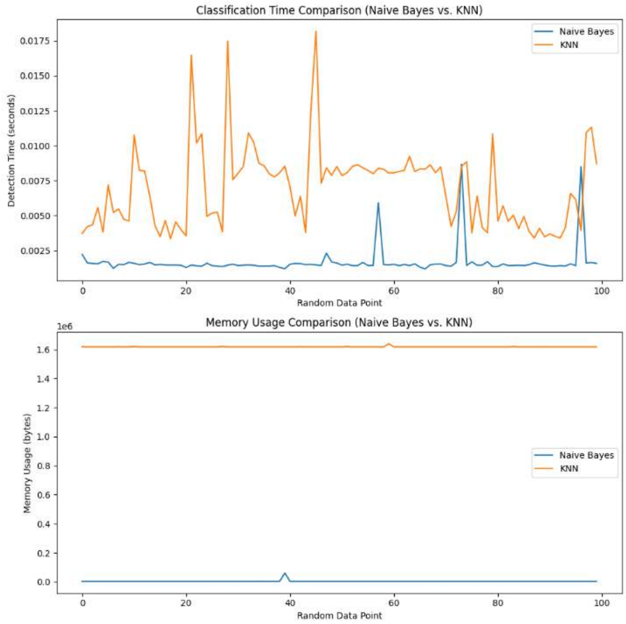

Fig. 5. Algorithm classification speed and memory usage comparison

TABLE III. ALGORITHM AVERAGE SPEED AND MEMORY USAGE COMPARISON

Algorithm	Average Classification Speed (seconds)	Average Memory Usage (bytes)
k-NN	0.007	1,618,185
Naive Bayes	0.0016	4028

The evaluation also showed the class that received most of the misclassifications was the "Less Fresh" class, likely because of its underrepresentation in the dataset. This imbalance likely caused the higher rate of errors associated with this class and affected the overall classification performance. The complete performance metrics of both algorithms are shown in Table IV.

TABLE IV. MODEL PERFORMANCE COMPARISON.

Algorithm	Class	Precision	Recall	F1-score	Accuracy
k-NN	Fresh	1.00	1.00	1.00	1.00
	Less fresh	1.00	1.00	1.00	
	Reject	1.00	1.00	1.00	
	Average	1.00	1.00	1.00	
Naive Bayes	Fresh	0.99	1.00	1.00	0.99
	Less fresh	0.98	0.98	0.98	
	Reject	1.00	0.99	0.99	
	Average	0.99	0.99	0.99	

Consistent with previous research findings, our study confirmed that the k-NN algorithm consistently yields the highest accuracy rates among various classification methods. This aligns with existing literature highlighting k-NN's robustness in sensor data classification tasks.

In summary, while the k-NN model is effective in real-world applications where high classification accuracy is required, the Naïve Bayes model's higher classification speed and lower memory usage might also prove to be a suitable alternative, given the application's specific needs.

V. CONCLUSION

This study evaluated the performance of the k-Nearest Neighbors (k-NN) and Naive Bayes algorithms in detecting fish freshness using gas sensor data. Both models demonstrated strong potential for real-time, automated quality control in the seafood industry. The k-NN algorithm excelled in accuracy, with nearly flawless results, misclassifying only one sample out of 7,207. Although the Naive Bayes algorithm was slightly less precise, it still performed admirably, achieving a high accuracy rate while offering prediction times almost 7x faster and 400x less memory usage. These findings underscore the potential of combining machine learning algorithms with gas sensors to assess fish quality, presenting a promising approach to improving freshness detection and quality control systems in the seafood sector. However, the study's scope was somewhat limited, focusing on a single type of fish. Future research could expand on this by incorporating a wider variety of fish species to enhance model robustness and account for potential variations in quality. In conclusion, this research highlights the effectiveness of both k-NN and Naive Bayes algorithms in classifying fish freshness, offering valuable insights for developing real-time quality control solutions using machine learning and gas sensors.

REFERENCES

[1] Ministry of Marine Affairs and Fisheries. (2024). *Marine and Fisheries in Figures 2024* (Vol. 11). The Center for Data, Statistics, and Information, Indonesia.

[2] Food and Agriculture Organization (FAO). (2022). *The State of World Fisheries and Aquaculture*. [Online].

[3] Statista, "Fish consumption per capita in Indonesia 2014-2022," Statista, 2023. [Online]. Retrieved from https://www.statista.com/statistics/1225355/indonesia-fish-consumption-per-capita/.

[4] Jakarta Globe. (2023). Indonesia Sets $7.6b Fishery Export Target for 2023. Retrieved from https://jakartaglobe.id/business/indonesia-sets-76b-fishery-export-target-for-2023

[5] The Jakarta Post. (2024). Ministry blames outdated methods for disappointing 2023 fishery exports. Retrieved from https://www.thejakartapost.com/business/2024/01/11/ministry-blames-outdated-methods-for-disappointing-2023-fishery-exports.html

[6] Y. Wu, Y. Duan, Y. Wei, D. An, & J. Liu, "Application of intelligent and unmanned equipment in aquaculture: A review," *Computers and Electronics in Agriculture*, vol. 199, pp. 107201, 2022.

[7] R. Baldwin, & R. Freeman, "Risks and global supply chains: What we know and need to know." *Annual Review of Economics*, vol. 14, no. 1, pp. 153-180, 2022.

[8] H. Anwar, T. Anwar, & S. Murtaza, "Review on food quality assessment using machine learning and electronic nose system," *Biosensors and Bioelectronics: X*, vol. 14, pp. 100365, 2023.

[9] R. Saeed, H. Feng, X. Wang, X. Zhang, and Z. Fu, "Fish quality evaluation by sensor and machine learning: A mechanistic review," Food Control, vol. 138, p. 108902, 2022.

[10] L. Franceschelli, A. Berardinelli, S. Dabbou, L. Ragni, and M. Tartagni, "Sensing Technology for Fish Freshness and Safety: A Review," *Sensors*, vol. 21, no. 4, pp. 1373, 2021.

[11] D. R. Wijaya, N. F. Syarwan, M. A. Nugraha, D. Ananda, T. Fahrudin and R. Handayani, "Seafood Quality Detection Using Electronic Nose and Machine Learning Algorithms With Hyperparameter Optimization," *IEEE Access*, vol. 11, pp. 62484-62