JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 2, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i2.9357

# HYBRID INTELLIGENCE FOR PRECISION MILK QUALITY ASSESSMENT: INTEGRATING NAÏVE BAYES AND K-MEANS CLUSTERING

**Baik Budi[1*], Rizki Sarhans[2]**

[1,2] Department of Electrical Engineering, Universitas Andalas, Jalan Limau Manis, Kec. Pauh, Kota Padang, Sumatera Barat 25163; (0751) 72497

**Keywords:**
Milk Quality; Naïve Bayes; K-Means; Clustering; Classification.

**Correspondent Email:**
baikbudi@eng.unand.ac.id



Milk quality is a fundamental factor in the food industry, directly impacting consumer health and product market value. Conventional manual assessment is often inefficient for large-scale operations, requiring rapid, accurate automated solutions. This study proposes a hybrid machine learning approach integrating the Naïve Bayes algorithm for quality classification (Low, Medium, High) and K-Means Clustering for diagnostic analysis based on physicochemical similarities. Using a dataset of 1,059 samples with seven sensory attributes (pH, temperature, taste, odor, fat, turbidity, and color), the data were cleaned, normalized, and modeled. Evaluation results demonstrate that the Naïve Bayes method achieved an accuracy of 85.38%. Notably, the model achieved 100% precision in identifying low-quality milk, making it a highly reliable food safety filter. Concurrently, K-Means with k=5 (selected as the optimal value after testing k=2 - 7) successfully segmented the data into five distinct diagnostic clusters. Centroid analysis revealed that temperature and odor are the primary factors for distinguishing the root causes of quality degradation. This study concludes that combining classification and clustering methods significantly enhances quality control efficiency by providing both quality labels and diagnostic insights into the factors driving milk spoilage.

## 1. INTRODUCTION

Milk quality serves as a fundamental indicator of both consumption feasibility and economic value in the global livestock and food industries [1]. As a food substance with a highly complex nutritional matrix, milk provides a nearly perfect nutritional profile for human metabolism [2]. This comprehensive nutritional characteristic results in milk's categorization as a "nearly perfect food source." [3]. The dairy sector not only contributes to nutritional security but also serves as a backbone of rural economies worldwide [4], [5], [6]. However, this high nutrient density also renders milk highly perishable and vulnerable to the growth of pathogenic microorganisms [6]. The physical properties of milk, characterized by high moisture content and a near-neutral pH, facilitate the exponential proliferation of bacteria within a short time, necessitating immediate and accurate quality control [7], [8]

The degradation or spoilage of milk is frequently triggered by microbial contamination originating from the external environment during the production process [9]. These contaminating factors include the cow's skin condition, udder hygiene, the quality of water used for equipment sanitation, soil conditions around the farm, and airborne dust particles. Microbial activity rapidly alters the physicochemical composition of milk, leading to irreversible changes in pH, protein stability,

and fat oxidation. Without a strict, systematic quality control system, the high nutritional content of milk fails to deliver functional benefits, instead risking human health through foodborne diseases [10]. Product safety measures must always accompany efforts to increase milk availability, as the nutritional value of a food becomes irrelevant if it causes illness to consumers.

In the modern food industry, milk quality analysis is performed by measuring a series of physicochemical parameters, including acidity (pH), temperature, taste, aroma, fat content, turbidity, and colour [11]. Traditional methods rely heavily on titratable acidity, standard plate counts, and lactometer readings [12], [13]. Although accurate, conventional laboratory testing procedures currently face significant efficiency challenges. Manual examinations require certified experts, incur high operational costs, and entail lengthy analysis times, making them less than ideal for large-scale industrial applications that require rapid decision-making. Consequently, the implementation of machine learning-based "Smart System" technology has emerged as a potential solution to automatically, consistently, and in real time classify milk quality.

A review of previous research indicates that the development of algorithms for food quality classification has been a primary focus over the last decade [14], [15]. Research has used Support Vector Machines (SVMs) for cow milk classification based on pH and temperature parameters, achieving good accuracy but noting a high computational load on large milk datasets. Conversely, the implementation of K-Nearest Neighbour (KNN) provided rapid results but proved highly sensitive to noise in sensor data, which is common in farm environments. Additionally, the use of Artificial Neural Networks (ANN) successfully handled non-linear data patterns for milk adulteration detection. However, the model remains difficult for field practitioners to interpret directly due to its "black box" nature.

To address the limitations of high computational costs and the lack of transparency in decision-making, hybrid models are emerging [16], [17]. A fundamental study by Budi et al. [18], [19], [20] demonstrated that a hybrid integration of Naive Bayes and K-Means Clustering can effectively handle complex environmental data by combining the predictive power of supervised learning with the structural insights of unsupervised clustering. Their work on air quality assessment provided a more robust framework for multidimensional datasets by using K-Means to identify data patterns before probabilistic **classification.** This methodological breakthrough suggests that combining supervised and unsupervised learning can mitigate the impact of noisy data and improve the reliability of smart monitoring systems across various agricultural domains.

Despite these advancements, a significant **research gap** persists in applying machine learning to dairy quality control. Most single-algorithm models struggle with "class imbalance"—a common issue in raw milk datasets, where "medium" quality samples often dominate the distribution, leading the model to be biased against "low" or "high" quality samples [19], [20]. Furthermore, supervised models often cannot autonomously validate whether predefined quality labels align with the natural, objective distribution of the physicochemical data [21].

The **novelty** of this research lies in the adaptation of the hybrid intelligence framework established in to the specific domain of dairy science. By synergizing the Naive Bayes algorithm with K-Means Clustering, this study introduces a dual-layer validation mechanism. Naive Bayes offers unparalleled processing speed and memory efficiency for Edge AI. At the same time, K-Means Clustering provides an unsupervised layer for identifying natural groupings in features without relying on initial human labels. This hybrid approach ensures that the classification boundaries reflect the natural clusters in milk.

Based on the background and identified gaps, the **objectives** of this research are: (1) To develop a robust classification model for milk quality (categorized into Low, Medium, and High) using the Naive Bayes algorithm to ensure high-speed decision-making, (2) To analyze the natural structural distribution of milk physicochemical parameters through K-Means Clustering to identify hidden patterns in milk degradation, and (3) To evaluate the performance and consistency of the integrated hybrid model in providing reliable quality assessments for the dairy processing industry.

The results of this research are projected to serve as a foundation for the development of smart sensor-based milk quality monitoring devices applicable to both local farmers and large-scale industries. By providing a computationally efficient yet highly accurate model, this study contributes to the digital transformation of the dairy industry, ensuring food safety and economic sustainability.

## 2. LITERATURE REVIEW

The Naïve Bayes technique is a probabilistic machine learning model based on Bayes' Theorem that calculates the likelihood of an event occurring given prior conditions. In the context of milk quality assessment, this algorithm predicts the quality class (H) based on a sensory attribute vector (X) comprising pH, temperature, and fat content. A fundamental characteristic of this algorithm is the "naïve" assumption that each attribute in the dataset is conditionally independent of the others, given the class label.

The mathematical formulation of Bayes' Theorem is expressed in Equation (1):

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Where:

- *X:* Data point with an unknown class (feature vector).
- *H*: Hypothesis that the data point belongs to a specific class.
- *P(H|X)*: Posterior probability; the probability of hypothesis *H* given condition *X*.
- *P(H)*: Prior probability; the initial probability of hypothesis *H*.
- *P(X|H)*: Likelihood; the probability of observing data *X* given hypothesis *H*.
- *P(X)*: Evidence; the total probability of observing data *X*.

The classification process selects the class with the highest posterior probability. The application of Naïve Bayes to data classification has been shown to offer superior computational efficiency, particularly in real-time analytical applications.

K-Means is a prominent unsupervised learning algorithm that partitions data into *k* distinct clusters based on the similarity of physicochemical characteristics, without using predefined quality labels. Its primary objective is to minimize intra-cluster variance (similarities within a group) while maximizing inter-cluster variance (differences between groups). This algorithm is widely utilized for data segmentation due to its simplicity, speed, and scalability.

The procedural steps for the K-Means algorithm implemented in this study are as follows:

1. Initialization: Select *k* initial cluster centers (centroids) randomly from the milk dataset. This research uses k=5 to achieve a more granular segmentation of milk characteristics, enabling the detection of subtle quality deviations.
2. Distance Calculation: Compute the distance between each milk sample and every centroid using the **Euclidean Distance** formula, as shown in Equation (2) below:

$$d(x,\mu) = \sqrt{\sum_{i-1}^{n}(x_i - \mu_j)^2} \quad (2)$$

Where:

- $x_i$: Criterion data (physicochemical attributes).
- $x_j$: Centroid of the $j-th$ cluster.

3. Assignment: Assign each milk sample to the cluster with the nearest centroid based on the calculated distance.
4. Centroid Update: Recalculate the position of the new cluster centers by taking the arithmetic mean of all data points currently assigned to that cluster.
5. Iteration: Repeat steps 2 through 4 until the centroid positions remain stable (convergence) or a predefined maximum number of iterations is reached.

The advantages of implementing K-Means Clustering in this hybrid model include:

1. Algorithmic Simplicity: High execution speed and ease of implementation.
2. Interpretability: The results are easy to visualize and understand in the context of physicochemical groupings.
3. Pattern Discovery: It yields superior results in identifying natural data structures when the dataset contains distinct, well-separated distributions.

## 3. RESEARCH METHODOLOGY

This research follows a structured computational framework for classifying milk

quality by integrating supervised and unsupervised learning techniques. The entire process, from data acquisition to performance evaluation, is illustrated in the flowchart in Figure 1 below:

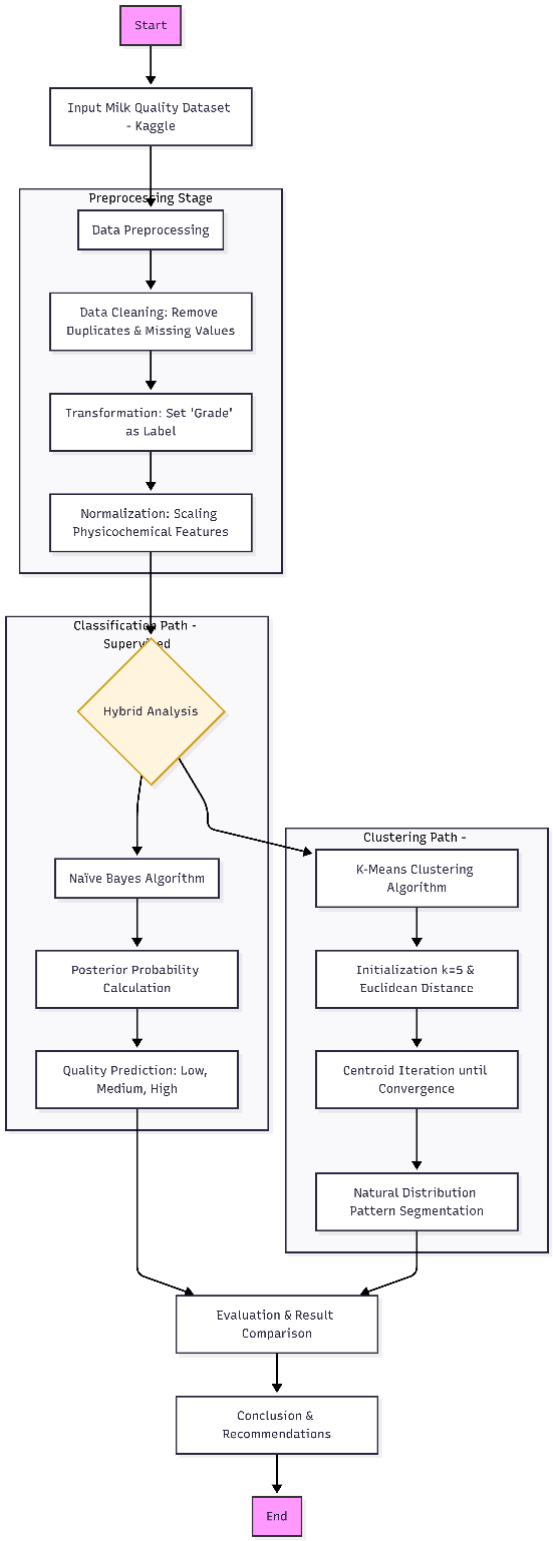


Figure 1. Research Methodology Flowchart

### 3.1 Data Acquisition and Preprocessing

The fundamental stage before predictive modeling involves rigorous data preprocessing to ensure data quality and computational efficiency. This study utilizes a raw dataset sourced from Kaggle [22] containing physicochemical parameters of raw milk (pH, Temperature, Taste, Odor, Fat, Turbidity, and Color). Data can be shown in Table 1 below:

Table 1. Attribute from the dataset Milk Quality

| No | Attribute | Description | Data Type |
|---|---|---|---|
| 1 | pH | Acidity/alkalinity level of milk (Range: 3 - 9.5) | Numeric |
| 2 | Temperature | Milk temperature during measurement (Storage indicator) | Numeric |
| 3 | Taste | Taste indicator (1: Good, 0: Bad) | Binary |
| 4 | Odor | Smell indicator (1: Fresh, 0: Not fresh) | Binary |
| 5 | Fat | Fat content in the milk | Numeric |
| 6 | Turbidity | Level of milk cloudiness (Purity indicator) | Binary |
| 7 | Colour | Milk color (Color spectrum value range) | Numeric |
| 8 | Target | Quality Class (Low, Medium, High) | Nominal |

The raw data is processed using specialized data analysis software through the following systematic stages:

1. **Data Cleaning:** This process involves removing duplicate records and data points with missing values to prevent bias. The **Replace Missing Values** operator handles null values, ensuring the dataset is complete.
2. **Data Transformation/Set Role):** To define the classification task clearly, the role of each attribute is specified. The 'Grade' column is designated as the **target (label)** attribute for predictive modeling, while all other physicochemical parameters are treated as regular features.

3. **Normalization:** The dataset is scaled using the **Normalize operator**. This step is vital for distance-based algorithms like K-Means Clustering, which are highly sensitive to Euclidean distance metrics. Normalization ensures that attributes with naturally larger numerical ranges (such as Temperature) do not dominate attributes with smaller ranges (such as pH), allowing each feature to contribute equally to the distance calculation.

### 3.2 Algorithmic Implementation

This study adopts and adapts the hybrid intelligence framework established by Budi et al [18], which has been proven effective in environmental data analysis. This research uses data downloaded from Kaggle.com [22]. The integration of Naive Bayes and K-Means serves two distinct purposes:

1. K-Means Clustering (Unsupervised Layer): Used to identify natural structural groupings within the milk samples. By setting k 1 until 7, the algorithm explores subtle variations in milk characteristics that might not be captured by broad "Low/Medium/High" labels.
2. Naïve Bayes (Supervised Layer): Once the data structure is understood, Naive Bayes is employed for rapid, probabilistic classification. This algorithm computes the posterior probability P(H|X) for assigning each sample to a quality class.

### 3.3 Performance Evaluation Layer: Confusion Matrix Analysis

To validate the robustness of the proposed hybrid model, a **Confusion Matrix** analysis is conducted to provide detailed performance metrics. While simple accuracy provides an overall correctness metric, it is insufficient for evaluating a multi-class classification model, especially when class imbalance is present. The Confusion Matrix provides a detailed visualization of correct predictions (True Positives and True Negatives) and misclassifications (False Positives and False Negatives) for each milk quality class: Low, Medium, and High.

The key performance metrics derived from this matrix are critical for ensuring food safety:

- Accuracy: Overall precision of the prediction.
- Precision: Model's ability to not mislabel unsafe milk as safe.
- Recall (Sensitivity): Crucial metric ensuring that *all* unsafe (low-quality) milk samples are correctly identified.
- F1-Score: The harmonic mean provides a balance between precision and recall.

## 4. RESULTS AND DISCUSSION

### 4.1 Quantitative Analysis of Naïve Bayes Performance

The experimental results, validated through an 80-20 data split, demonstrate that the Naïve Bayes classifier provides a robust framework for milk quality assessment. While the overall **Accuracy of 85.38%** is high, the model's true strength lies in its class-specific performance, as detailed in Table 2 below.

**Table 2. Confusion Matrix of Naïve Bayes**

| Actual | High | Low | Medium | Total |
|---|---|---|---|---|
| High | 50 | 0 | 1 | 51 |
| Low | 4 | 80 | 2 | 86 |
| Medium | 24 | 0 | 51 | 75 |
| Total | 78 | 80 | 54 | 212 |

1. **Reliability in Extreme Quality Detection (Safety Criticality)**

The model achieved a Recall of 0.98 for High and 0.93 for Low grades. From a food safety perspective, the high recall for the "Low" class is the most critical metric. It indicates a very low rate of False Negatives (spoiled milk being labeled as good). Statistically, Naïve Bayes excels here because the physicochemical profiles of "Low" quality milk (e.g., pH < 6.0 or Temperature > 15°C) are distinct and distance-separated from the "High" quality cluster, allowing the probabilistic model to draw clear decision boundaries.

2. **Economic Efficiency through Precision**

A standout finding is the Precision of 1.00 for the Low class. In industrial operations, a Precision of 1.00 means there are zero False Positives for spoiled milk. This has direct economic implications: the system never misclassifies high-quality milk as "Low." Consequently, producers can avoid "Type I Error" costs—preventing the unnecessary disposal of viable products —a common inefficiency in manual inspection.

3. **The "Medium-High" Overlap: A Biochemical Bias**

The primary source of accuracy loss is the misclassification of **24 Medium-grade samples as High**. This is not merely an algorithmic failure but a reflection of **Biochemical Continuity**. Milk degradation is a continuous process, not a discrete one. The transition from "High" (pH ~6.7) to "Medium" (pH ~6.5) is subtle. Because Naïve Bayes assumes feature independence, it may struggle when two classes exhibit high feature overlap and high feature variance [23].

### 4.2 K-Means Clustering Analysis

The exploration of natural data distribution patterns was conducted using the K-Means Clustering algorithm. This unsupervised learning method partitions data based on similarities in physicochemical characteristics, without relying on predefined class labels [24]. This process was executed using Altair AI Studio, following a workflow that began with handling missing values via the *Replace Missing Values* operator, defining attribute roles (*Set Role*), and a data normalization phase. The use of the *Normalize* operator is critical in distance-based algorithms to equalize feature scales, ensuring that attributes with large value ranges—such as temperature—do not disproportionately influence distance calculations compared to those with smaller ranges, such as pH. The centroid values for each cluster can be shown in the Table. 3 Below:

Table. 3 Centroid value for each cluster

| Attribute | Clu_0 | Clu_1 | Clu_2 | Clu_3 | Clu_4 |
|---|---|---|---|---|---|
| **pH** | -0.839 | -0.243 | 1.030 | 0.070 | 0.796 |
| **Temp** | -0.156 | -0.556 | 2.148 | -0.067 | -0.223 |
| **Taste** | -0.119 | -0.245 | -0.444 | 0.725 | 0.237 |
| **Odor** | 1.145 | -0.736 | -0.231 | 0.368 | -0.320 |
| **Fat** | 0.699 | -0.324 | -0.159 | 0.112 | -0.311 |
| **Turbidity** | 1.009 | -0.962 | 0.274 | -0.717 | 1.018 |
| **Colour** | 0.524 | 0.498 | 0.377 | -1.549 | -0.405 |

The K-Means algorithm was applied with *k=5* and the Bregman Divergence distance measure. Within the context of Altair AI Studio, this measure is equivalent to the distance approach in a standardized feature space, providing robustness against variations in data distribution. Cluster centers (centroids) were automatically initialized to provide an optimal starting point, and the iterative process continued until convergence. Based on processing results for 1,059 items, the cluster distribution revealed a clearly segmented pattern, as shown in the Table. 4 below:

Table 4. Cluster result from K-Means with k=5

| Cluster | Distribution Amount |
|---|---|
| Cluster 0 | 238 Items |
| Cluster 1 | 339 Items (largest) |
| Cluster 2 | 129 Items (smallest) |
| Cluster 3 | 174 Items |
| Cluster 4 | 179 Items |
| Total | 1059 Items |

The centroid table maps the normalized central values for each cluster to each attribute. Positive values indicate that the attribute is above the global average, while negative values indicate that it is below the average. A sharpened interpretation of each cluster's characteristics based on the processed data is as follows:

1. Cluster 0 (High-Risk Milk - Odor & Turbidity): This cluster is characterized by extremely high values for Odor (1.145) and Turbidity (1.009). High levels of turbidity and pungent aromas empirically correlate with microbial activity degrading milk proteins, indicating a Low quality category.
2. Cluster 2 (Temperature & pH Anomalies): This group exhibits significantly high Temperature (2.148) and a tendency toward an alkaline pH (1.030). High temperatures accelerate chemical and biological reaction rates, which, in food safety standards, signifies milk that has been exposed to excessive heat and is highly susceptible to spoilage.
3. Cluster 3 (Specific Taste & Color Profile): This cluster stands out with the highest Taste (0.725) value but the lowest Color (-1.549) value. These characteristics reflect milk with a strong flavor profile and a very specific color, often associated with High or premium Medium quality.
4. Cluster 4 (Physical Anomaly Without Odor): This group displays an alkaline pH (0.766) and high Turbidity (1.018), yet maintains a low Odor (-0.320) value. This represents milk samples undergoing anomalous physical texture changes without yet exhibiting aromatic decay.

5. Cluster 1 (Standard Characteristics): All centroid values in this cluster are near zero or negative, indicating a group of samples with neutral and stable physicochemical features.

The integration of K-Means provides deeper diagnostic insights than a standalone classification method like Naïve Bayes. Through this hybrid approach, the system does not merely assign a quality label; it uncovers the technical factors behind the differences—whether caused by extreme temperatures, shifts in aroma, or pH imbalances.

## 5. CONCLUSION

1. The implementation of the Naïve Bayes algorithm as a supervised learning model achieved a significant overall accuracy of 85.38%. The model demonstrated superior performance in detecting extreme quality categories, with sensitivities (recalls) of 0.98 for the High class and 0.93 for the Low class. A critical finding is the 1.00 (100%) precision achieved for the Low-quality category, ensuring the system is entirely reliable in identifying spoiled milk without the risk of false positives. In an industrial context, this absolute precision prevents the unnecessary disposal of viable products while maintaining a rigorous food safety filter.
2. The K-Means clustering analysis, optimized at k=5, revealed natural data distribution patterns that a standalone classification model could not capture. Processing 1,059 items yielded deep diagnostic insights through centroid analysis. Temperature and Odor were identified as the primary differentiators among the data groups. Specifically, Cluster 2 successfully isolated milk samples exposed to extreme temperatures (2.148 standard deviations above average), while Cluster 0 identified physical degradation characterized by high odor (1.145) and turbidity (1.009). This capability allows the system not only to label quality but also to provide a technical justification for spoilage, such as storage mismanagement or bacterial contamination.
3. The synergy between these two methods transforms the system from a simple labeling tool into a comprehensive diagnostic framework. In this architecture, Naïve Bayes serves as the front-end for rapid sorting (Good/Bad). At the same time, K-Means acts as a back-end laboratory assistant to trace the root causes of quality degradation.

## REFFERENCES


[1] H. Priyashantha, "World dairy system sustainability: a milk quality perspective," *Frontiers in Sustainable Resource Management*, vol. 4, Jun. 2025, doi: 10.3389/fsrma.2025.1572962.

[2] A. Moghaddamjoo and M. Allam, "Techniques in Array Processing by Means of Transformations," *Control and dynamic systems*, vol. 69, pp. 133–180, 1995.

[3] W.-K. Chen, *Linear Networks and Systems: Algorithms and Computer-Aided Implementations*. Belmont, CA: Wadsworth Publishing Company, 1993.

[4] M. Siddiky, "Dairying in South Asian region: opportunities, challenges and way forward," *SAARC Journal of Agriculture*, vol. 15, p. 173, Jul. 2017, doi: 10.3329/sja.v15i1.33164.

[5] V. Nimbalkar, H. K. Verma, and J. Singh, "Dairy Farming Innovations for Productivity Enhancement," in *New Advances in the Dairy Industry*, M. S. Qureshi, Ed. London: IntechOpen, 2021, ch. 5, doi: 10.5772/intechopen.101373.

[6] A. A. Gabriel *et al.*, "Fates of pathogenic bacteria in time-temperature-abused and Holder-pasteurized human donor-, infant formula-, and full cream cow's milk," *Food Microbiology*, vol. 89, p. 103450, Aug. 2020, doi: 10.1016/j.fm.2020.103450.

[7] L. Bass, P. Clements, and R. Kazman, "Software Architecture in Practice 2nd Edition," Jan. 2003.

[8] T. J. van Weert and R. K. Munro, Eds., *Informatics and the Digital Society: Social, Ethical and Cognitive Issues*, vol. 121. Boston: Springer Science & Business Media, 2003, doi: 10.1007/978-0-387-35663-1.

[9] M. W. Dixon, "Application of neural networks to solve the routing problem in communication networks," Doctoral Thesis, Div. Sci. Eng., Murdoch Univ., Perth, Australia, 2004.

[10] D. Putri, G. Forda Nama, and W. Sulistiono, "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier," ***Jurnal Informatika dan Teknik Elektro Terapan***, vol. 10, Jan. 2022, doi: 10.23960/jitet.v10i1.2262.

[11] A. Ismanto, F. Ardhani, and M. Marhamah, "The Effect of Traditional Transportation Using Cool Box on Quality of Fresh Milk and Frozen Milk from Peternakan Sapi Terpadu Sangatta to Samarinda East Kalimantan," *Buletin Peternakan*, vol. 42, no. 3, pp. 241–245, Aug. 2018, doi: 10.21059/buletinpeternak.v42i3.31559.

[12] C. Guajardo *et al.*, "MILK QUALITY AND DAIRY PRODUCT DEVELOPMENT OF A NORMANDE COW HERD IN THE REGION OF ÑUBLE, CHILE," *Chilean journal of agricultural & animal sciences*, vol. 36, pp. 190–197, Dec. 2020, doi: 10.29393/CHJAAS36-17MQCG80017.

[13] G. Wanjala, "Microbiological quality and safety of raw and pasteurized milk marketed in and around Nairobi region," *AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPEMENT*, vol. 17, pp. 11518–11532, Mar. 2017, doi: 10.18697/ajfand.77.15320.

[14] J. Chauvin *et al.*, "Advanced Optical Technologies in Food Quality and Waste Management," in *Innovation in the Food Sector Through the Valorization of Food and Agro-Food By-Products*, A. N. de Barros and I. Gouvinhas, Eds. London: IntechOpen, 2021, ch. 6, doi: 10.5772/intechopen.97624.

[15] K. Chhetri, "Applications of Artificial Intelligence and Machine Learning in Food Quality Control and Safety Assessment," *Food Engineering Reviews*, vol. 16, pp. 1–21, Dec. 2023, doi: 10.1007/s12393-023-09363-1.

[16] A. Siddique *et al.*, "Big data analytics in food industry: a state-of-the-art literature review," *npj Science of Food*, vol. 9, no. 1, p. 36, Mar. 2025, doi: 10.1038/s41538-025-00394-y.

[17] S. Wolfert *et al.*, "Navigating the Twilight Zone: Pathways towards digital transformation of food systems," Sep. 2021, doi: 10.18174/552346.

[18] B. Budi, F. A. Andhika, and T. Mahardika, "PENILAIAN KUALITAS UDARA DAN ANALISIS POLUSI BERBASIS ALGORITMA NAIVE BAYES DAN KLUSTRERISASI DATA DENGAN K-MEANS," *Jurnal Informatika dan Teknik Elektro Terapan* ***(JITET)***, vol. 13, no. 3S1, pp. 408–415, 2025, doi: 10.23960/jitet.v13i3S1.7630.

[19] M. Khan, V. Thorup, and Z. Luo, "Delineating Mastitis Cases in Dairy Cows: Development of an IoT-Enabled Intelligent Decision Support System for Dairy Farms," *IEEE Transactions on Industrial Informatics*, vol. PP, Apr. 2024, doi: 10.1109/TII.2024.3384594.

[20] O. Kashongwe *et al.*, "Influence of Preprocessing Methods of Automated Milking Systems Data on Prediction of Mastitis with Machine Learning Models," *AgriEngineering*, vol. 6, no. 3, pp. 3427–3442, 2024, doi: 10.3390/agriengineering6030195.

[21] G. Vishwakarma, A. Sonpal, and J. Hachmann, "Metrics for benchmarking and uncertainty quantification: Quality, applicability, and a path to best practices for machine learning in chemistry," *Trends in Chemistry*, vol. 3, no. 2, pp. 146–156, 2021, doi: 10.1016/j.trechm.2020.12.004.

[22] C. Shrijayan, "Milk Quality Prediction Dataset," *Kaggle*, 2024. [Online]. Available: https://www.kaggle.com/datasets/cpluzshrijayan/milkquality

[23] F. Ramadhani, Al-Khowarizmi, and I. P. Sari, "Improving the Performance of Naïve Bayes Algorithm by Reducing the Attributes of Dataset Using Gain Ratio and Adaboost," in *2021 International Conference on Computer Science and Engineering (IC2SE)*, 2021, vol. 1, pp. 1–5, doi: 10.1109/IC2SE52832.2021.9792027.

[24] A. Karahoca, *Data Mining Applications in Engineering and Medicine*. London: IntechOpen, 2012, doi: 10.5772/2616.