Research Article

Computer Vision-Based Rapid Classification of Raw Milk Quality in Indonesian Smallholder Dairy Systems

Ahmad Khoirul Umam[1,4], Lilik Eka Radiati[1]*, Tri Eko Susilorini[1], Fitri Utaminingrum[2], Herly Evanuarini[1], Agus Susilo[1], Puguh Surjowardojo[1], Dimas Firmanda Al Riza[3], Muhammad Helmi[4], Abiyyu Rafi Amrullah[4]

[1]Faculty of Animal Science and Technology, Universitas Brawijaya, Malang, Indonesia; [2]Faculty of Computer Science, Universitas Brawijaya, Malang, Indonesia; [3]Faculty of Agroindustrial Technology and Biosystems, Universitas Brawijaya, Malang, Indonesia; [4]Faculty of Animal Science and Technology, PSDKU Universitas Brawijaya, Kediri, Indonesia.

Abstract | Deterioration of raw milk quality during tropical ambient storage (25–30°C) poses critical food safety challenges for Indonesian smallholder dairy cooperatives. Conventional laboratory testing (24–48 hours) precludes real time quality gatekeeping at the point of collection . This study developed and validated a computer vision system integrating Gray Level Co-occurrence Matrix (GLCM) texture analysis with machine learning classifiers for rapid, cost-effective milk quality screening. Three hundred raw milk samples from smallholder cooperative farms in Malang District, East Java, Indonesia, were classified as Good (n = 135, 45.0%), Abnormal (n = 108, 36.0%), or Defective (n = 57, 19.0%) per SNI 3141.1:2011 standards using Total Plate Count (TPC), pH, and titratable acidity. Sixteen image features (4 GLCM texture + 6 RGB statistics + 6 HSV statistics) were extracted from standardised digital images and evaluated using K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), and XGBoost classifiers with five-fold cross-validation. XGBoost achieved the highest classification accuracy (93.2%), significantly outperforming SVM (91.7%, P = 0.023) and K-NN (87.6%, P < 0.001) at 2.1 seconds per sample. GLCM features dominated classification importance (88.8%), led by red-channel Contrast (55.6%) and Homogeneity (33.2%), showing strong biological alignment with TPC (r = 0.892) and titratable acidity (r = 0.831), reflecting GLCM sensitivity to bacterial proliferation and protein degradation rather than serving as independent validation . In the test dataset, zero Defective samples were misclassified as Good; three were conservatively misclassified as Abnormal maintaining rejection status but representing classification level under severity rather than a food safety failure. Temporal analysis identified a 12-hour critical quality window before TPC crossed the SNI threshold (1×10^6 CFU/mL) under tropical ambient storage. The XGBoost–GLCM system provides a validated, real-time preliminary screening tool that reduces collection-point decision time by >99.9% relative to conventional laboratory testing, intended to complement and not for replace confirmatory methods, supporting Indonesia's national dairy self-sufficiency and public health protection objectives.

Keywords | Computer vision, Machine learning, GLCM texture analysis, K-NN, SVM, XGBoost

Received | February 23, 2026; **Accepted** | April 05, 2026; **Published** | June 02, 2026

***Correspondence** | Lilik Eka Radiati, Department of Animal Products Technology, Faculty of Animal Science and Technology, Universitas Brawijaya, Malang, Indonesia; **Email:** lilik.eka@ub.ac.id

Citation | Umam AK, Radiati LE, Susilorini TE, Utaminingrum F, Evanuarini H, Susilo A, Surjowardojo P, Al-Riza DF, Helmi M, Amrullah AR (2026). Computer vision-based rapid classification of raw milk quality in Indonesian smallholder dairy systems. Adv. Anim. Vet. Sci., 14(6):1190-1205.
DOI | https://dx.doi.org/10.17582/journal.aavs/2026/14.6.1190.1205
ISSN (Online) | 2307-8316

INTRODUCTION

Food safety in dairy production constitutes a considerable public health concern globally. Unpasteurised milk carries an estimated 150-fold higher risk of foodborne disease outbreaks compared to pasteurised milk (Mungai *et al*., 2015), with *Salmonella* spp., *Escherichia coli* O157:H7, *Listeria monocytogenes*, and *Campylobacter jejuni* as the primary pathogens of concern. In developing tropical countries, where cold-chain infrastructure is limited and raw milk consumption by rural households is common, this burden is amplified: 38–52% of farm-collected raw milk in Southeast Asian smallholder systems fails microbiological safety thresholds before reaching processing facilities (Jaenudin *et al*., 2017). In Indonesia, these risks are compounded by tropical ambient temperatures of 25–30°C that accelerate bacterial proliferation to unsafe levels within hours of milking, while approximately 95% of dairy farmers operate with fewer than five cows and without cold storage infrastructure (Fadillah *et al*., 2023). Despite annual production of 1 million metric tons, the sector meets only 22% of national demand (Direktorat Jenderal Peternakan dan Kesehatan Hewan, 2023) with quality failures identified as the primary constraint to expansion.

Conventional milk quality assessments, including Total Plate Count (TPC), pH, and titratable acidity, require 24 to 72 hours and specialized laboratory facilities (Bintsis, 2017). This constitutes a major challenge in Indonesia's primary dairy regions East Java (45% of production), Central Java (30%), and West Java (20%) where approximately 15,000 farmers in Malang district rely on only two or three laboratories. Testing expenses range from IDR 50,000 to 100,000 per sample, representing 5-10% of the daily milk value; consequently, farmers conduct tests weekly or monthly rather than for every batch. Tropical temperatures (25-30°C) reduce the safe handling period of milk to 12-18 hours post-milking, compared to 24-36 hours in cooler climates. Furthermore, over 90% of Indonesian herds consist of heat-stressed Holstein-Friesians, which produce milk with elevated somatic cell counts that deteriorate rapidly even in the absence of mastitis (Das *et al*., 2016). Farmers generate daily incomes of IDR 180,000-210,000 (USD 12-14) from 30 liters sold at IDR 7,000-8,000 per liter, constituting nearly half of a household's monthly income of IDR 2.5-3 million. Therefore, milk rejection results in notable financial losses and perpetuates vulnerability among rural dairy families (Young *et al*., 2026).

Computer vision and artificial intelligence technologies offer validated solutions for rapid, non-destructive food quality assessment across diverse matrices (Bhargava and Bansal, 2021). For liquid and semi-liquid food products specifically, image-based texture analysis has demonstrated strong discriminative performance: GLCM-derived features have been applied to milk powder surface characterisation (Ding *et al*., 2022a), colour-texture extraction has predicted biochemical quality indicators in liquid food matrices (Menozzi *et al*., 2025; Saleh and Lee, 2023), and near-infrared imaging combined with spectral features has classified milk by geographical origin (Wu *et al*., 2024). However, direct application of GLCM texture analysis to raw liquid milk quality classification under tropical ambient storage conditions the operationally critical scenario for Indonesian smallholder cooperatives has not been systematically validated. GLCM quantifies spatial pixel intensity relationships that reflect microstructural changes during milk deterioration, including bacterial proliferation, increasing turbidity, enzymatic protein degradation, casein micelle disruption, and fat globule coalescence (Haralick *et al*., 1973; Brink *et al*., 2020; Umam *et al*., 2025), producing descriptors Contrast, Homogeneity, Energy, and Correlation that characterise progressive quality degradation (Menozzi *et al*., 2025). Machine learning classifiers (K-NN, SVM, XGBoost) trained on such features show strong food quality classification performance suitable for resource-limited deployment (Sarker, 2021), but their application to this specific context requires empirical validation.

Real-time quality assessment at the farm or cooperative collection level facilitates prompt interventions such as immediate delivery, emergency cooling, or diversion to alternative uses, thereby preventing contaminated milk from entering processing chains (Weldeabezgi *et al*., 2020). From a One Health perspective, real-time quality screening at the cooperative level contributes to food safety (preventing contaminated milk from entering processing chains) and public health (reducing foodborne illness burden) simultaneously (Destoumieux-Garzón *et al*., 2018). While the system classifies milk quality based on composite deterioration markers TPC, pH, and titratable acidity rather than mastitis-specific biomarkers, systematic quality deviations detected at collection may prompt veterinary follow-up at the farm level, providing an indirect signal relevant to herd health monitoring. The system cannot, however, distinguish milk from a cow with subclinical mastitis from milk that has undergone equivalent spoilage through prolonged ambient storage; direct mastitis detection would require somatic cell count-specific imaging or dedicated biomarker assays beyond the scope of this study.In Indonesia, improved quality monitoring directly supports the 2025-2045 Food Security Roadmap, which aims to achieve 40% dairy self-sufficiency through increased smallholder productivity and quality improvements (Government of the Republic of Indonesia, 2024). Given climate change projections of rising tropical temperatures and rainfall variability, rapid quality-assessment technologies are increasingly essential

as deterioration kinetics accelerate unpredictably (Godde *et al*., 2021; Thornton *et al*., 2022).

This study developed and validated a computer vision system integrating GLCM texture analysis with K-NN, SVM, and XGBoost classifiers for three-class raw milk quality classification (Good, Abnormal, Defective) under tropical ambient storage conditions representative of Indonesian smallholder cooperatives. By providing the first systematic validation of GLCM-based machine learning for this operational context, this research addresses a critical gap in food safety technology for resource-limited tropical dairy systems.

MATERIALS AND METHODS

Study design and sample collection

This study was conducted at smallholder dairy farms in Malang, East Java, Indonesia, which represent the predominant production system in the Indonesian dairy sector, where approximately 95% of dairy farmers operate with fewer than five cows per household. Five representative smallholder farms were selected, each maintaining between one and eight Holstein-Friesian cattle under semi-intensive production systems typical of East Java dairy operations. A total of 20 cows were sampled across the five farms, generating 20 independent baseline milk batches (4 batches per farm on average). Milk samples were collected over a three-month period (November 2025 to January 2026) to capture variations in milk quality under smallholder management practices. Each cow contributed one independent milk batch, defined as the total morning milking yield from a single animal collected on a single occasion milk was not pooled across cows within or across farms. The 20 batches therefore represent 20 biologically independent sampling units, with within-farm variability captured through the multiple individual-cow batches collected per farm

Fresh milk samples were obtained from healthy cows during morning milking sessions using sterile collection procedures. Each cow's udder health status was verified by a veterinarian prior to collection. To investigate milk quality deterioration under realistic tropical storage conditions faced by smallholder farmers without cold chain infrastructure, fresh milk samples whose baseline quality was confirmed through initial laboratory testing were stored at ambient temperatures of 25 to 30 °C, typical of Indonesian room temperature. Temperature was not artificially controlled at a fixed point, as the study objective was to simulate realistic field deterioration trajectories rather than controlled incubation. It is acknowledged that bacterial doubling times differ approximately two-fold between 25°C and 30°C; the 12-hour critical quality window identified in this study therefore represents an estimate applicable to the observed mean storage temperature range and should be interpreted as a context-specific operational guideline rather than a fixed biological threshold

From each fresh milk batch, aliquots were sampled at nine time points: 0, 6, 12, 18, 24, 30, 36, 42, and 48 hours post-milking. Because multiple time points were drawn from each batch, statistical analyses accounted for the repeated-measures structure: cross-validation and McNemar's test were applied at the sample level, and batch-level independence was preserved by ensuring that no single batch contributed samples to both the training and test sets simultaneously A total of 300 milk samples were collected, comprising multiple time-point measurements from baseline fresh samples. From each batch at each time point, two aliquots were collected: one designated for microbiological analysis (TPC) and one for physicochemical analysis (pH and titratable acidity), yielding a maximum of 20 batches × 9 time points × 2 aliquots = 360 potential observations. After excluding aliquots from batches that failed baseline quality confirmation (n = 60 excluded across all time points), the final analytical dataset comprised 300 samples distributed as follows: 0 h (n = 34), 6 h (n = 34), 12 h (n = 33), 18 h (n = 33), 24 h (n = 34), 30 h (n = 33), 36 h (n = 33), 42 h (n = 33), and 48 h (n = 33). It is acknowledged that the two aliquots per batch per time point constitute analytical replicates sharing the same biological source; the statistical analysis accounted for this nesting structure by treating the batch-time-point unit as the primary sampling unit and applying batch-level train/test separation to prevent data leakage across model evaluation partitions.. This experimental design simulates the quality degradation trajectory from fresh milking to various delay scenarios before reaching cooperative collection points, reflecting real-world conditions in which smallholder farmers may experience transportation delays or limited access to immediate cooling facilities.

Quality classification and validation

Milk quality classification employed three laboratory reference methods at each storage time point: Total Plate Count (TPC), pH measurement, and titratable acidity determination. Samples were categorized into three quality classes according to the Indonesian National Standard SNI 3141.1:2011 (Badan Standardisasi Nasional, 2011), with class assignment contingent upon agreement across all three parameters. It is acknowledged that correlations subsequently reported between GLCM features and TPC or titratable acidity do not constitute independent external validation, as the same parameters were used to define ground truth labels. These correlations should be interpreted as evidence of biological interpretability

demonstrating that GLCM texture descriptors track the microstructural processes measured by the reference methods rather than as confirmation of model performance against an independent criterion. A truly independent validation would require ground truth defined by a method orthogonal to the classification parameters, such as sensory evaluation or flow cytometry; this is identified as a priority for future work.

- Good quality (n=135) refers to milk meeting SNI standards with TPC below 1×10^6 CFU/mL, pH between 6.5 and 6.7, and titratable acidity under 0.18%. These samples typically correspond to storage times of 0–12 hours. Classification was based exclusively on measured parameter values; storage time was not a classification criterion. As a descriptive observation, the majority of Good-quality samples were drawn from the 0–12 hour time points, consistent with the expected deterioration trajectory, but samples were not assigned to this class on the basis of time.
- Abnormal quality (n= 108): milk in which at least one parameter exceeded its SNI threshold TPC 1×10^6 to 5×10^6 CFU/mL, pH 6.3–6.5 or 6.7–7.0, and/or titratable acidity 0.18–0.25% while the remaining parameters had not yet reached Defective thresholds. As a descriptive observation, most Abnormal samples were drawn from the 18–30 hour time points, reflecting early-stage deterioration; however, class assignment was determined entirely by measured values.
- Defective quality (n= 57): milk in which multiple parameters showed severe deviation from SNI standards TPC > 5×10^6 CFU/mL, pH < 6.3 or > 7.0, and titratable acidity > 0.25% accompanied by marked discolouration, coagulation, and pronounced off-odours rendering milk unsuitable for human consumption. Most Defective samples were drawn from time points beyond 36 hours; however, as above, classification was based on measured parameters, not storage time.

The multi-parameter validation approach ensured robust quality classification by assigning samples to a class only when all three indicators (microbiological, chemical pH, and chemical acidity) consistently supported the classification. This triangulation method minimized misclassification risk and provided reliable ground truth labels for computer vision model training and validation.

Computer vision system development

Image acquisition

Image acquisition was performed using a custom-built imaging chamber (PRIME BOX) designed to provide controlled, reproducible lighting conditions suitable for deployment at cooperative collection points. The enclosure (20×20×30 cm) was equipped with white LED panels (5500K colour temperature, 2000 lux) for standard visible-light imaging. A 12-megapixel industrial camera (IDS UI-3370CP-C-HQ) was mounted at a fixed distance of 20 cm above the sample platform to ensure consistent image scale and perspective (Umam *et al*., 2025). This controlled laboratory setup was used to establish classification performance benchmarks under optimised imaging conditions. The authors acknowledge that this configuration requiring a custom enclosure, LED panels, and an industrial camera is not directly deployable in smallholder cooperative settings in its current form. Translation to a field-deployable system, potentially using smartphone cameras with a low-cost standardised enclosure, is identified as a priority for future development.

For each trial, 15 mL of milk was poured into a standardized transparent Petri dish (90 mm in diameter) and treated with methyl red dye. Methyl red dye (0.1% solution, 0.5 mL per sample) was added as a pH-sensitive colorimetric indicator to visually amplify the subtle hue shifts associated with acidification during bacterial growth, thereby enhancing the spectral contrast captured by the imaging system. The dye functions as a pH-sensitive colorimetric indicator, amplifying acidification-associated hue shifts for colour feature extraction (RGB/HSV channels); it does not directly influence GLCM texture features, which are computed from the grayscale intensity channel. This distinction was confirmed in preliminary trials (n= 30 samples per quality class, 90 total; paired dye-treated vs. untreated aliquots from the same batch): no significant difference was observed in any of the four GLCM texture metrics between conditions (paired t-test, all $P > 0.05$; minimum detectable effect size= 0.15 at $\alpha = 0.05$, power = 0.80). Colour features (RGB/HSV) showed expected dye-induced shifts, confirming that the dye's contribution is confined to the colour feature subset. The sample was allowed to settle for 3 minutes to eliminate surface disturbances. To capture comprehensive surface characteristics including color variations, textural patterns, and opacity changes indicative of quality deterioration four images were captured per sample at rotational angles of 0°, 90°, 180°, and 270°. All images were recorded at a resolution of 4096×3072 pixels in JPEG format to facilitate subsequent analysis

Image preprocessing and data augmentation

Images underwent standardized preprocessing pipeline followed the protocol of (Umam *et al*., 2025) including: (1) region of interest (ROI) extraction focusing on milk sample area with background removal using threshold-based segmentation, (2) conversion to grayscale for GLCM analysis while preserving RGB channels for color feature extraction, (3) resizing to 256×256 pixels for consistent feature computation, and (4) contrast enhancement using adaptive histogram equalization to amplify subtle textural differences.

To address class imbalance and improve classifier generalization, extensive data augmentation was applied to the training set. Augmentation techniques included: horizontal and vertical flipping, rotation (±15°), brightness adjustment (±20%), Gaussian noise addition (σ=0.01), and slight zoom variations (90-110%). Each original training sample generated 5 augmented variants; augmentation was applied exclusively to the training set (n= 1,050 original-sample feature vectors expanded to 6,300), while the validation set (n= 225) and independent test set (n= 225) comprised feature vectors extracted from non-augmented original images only. Classification performance reported on the independent test set therefore reflects model generalisability to unaugmented data rather than to augmented variants. The authors acknowledge that augmented feature distributions may not fully represent genuine biological variation encountered across different cooperatives, lighting environments, or operator handling an assumption that requires validation through prospective field testing. The reported 93.8% test-set accuracy should be interpreted as a laboratory-condition benchmark pending such validation.

GLCM feature extraction

Gray Level Co-occurrence Matrix (GLCM) analysis was performed to extract texture features from preprocessed milk images followed by Ding *et al*. (2022b). GLCM computes the spatial relationship between pixel pairs at specified distances and angles, generating a co-occurrence matrix that quantifies texture patterns. For each image, the GLCM was computed at four orientations (0°, 45°, 90°, 135°) with a pixel distance d=1 and averaged to obtain rotation-invariant features.

From each GLCM matrix, four Haralick texture descriptors were calculated:

- Contrast: measures local intensity variation, quantifying the degree of textural roughness (higher values indicate greater variation between neighboring pixels)
- Correlation: assesses linear dependency between pixel pairs, indicating texture directionality and pattern regularity
- Energy (Angular Second Moment): measures texture uniformity and homogeneity, with higher values indicating more ordered patterns
- Homogeneity (Inverse Difference Moment): quantifies closeness of GLCM element distribution to the diagonal, reflecting texture smoothness

Additionally, color-based features were extracted from RGB and HSV color spaces, including mean and standard deviation of each channel. The final feature vector for each sample consisted of 16 features: 4 GLCM texture features, 6 RGB statistics (mean and std for R, G, B), and 6 HSV statistics (mean and std for H, S, V). These features captured both textural changes associated with microbial growth and color alterations related to pH shifts and protein degradation.

Machine learning classification algorithms

Three supervised machine learning algorithms were selected to represent distinct classification paradigms applicable to small, structured feature datasets: K-NN as a non-parametric instance-based baseline, SVM as a margin-based kernel method suited to non-linear high-dimensional feature spaces, and XGBoost as an ensemble gradient boosting method with demonstrated robustness in structured food quality classification tasks. Neural network architectures were not included as the primary classifiers given the modest dataset size (300 original samples); deep learning approaches require substantially larger training sets for reliable generalisation. Random forest was evaluated in preliminary screening but showed performance within 0.8% of XGBoost accuracy with higher training overhead and was therefore not included in the primary comparison.

K-Nearest Neighbors (K-NN): A non-parametric algorithm that classifies samples based on majority voting among k nearest neighbors in feature space. Euclidean distance was used as the similarity metric. The optimal k value was determined through cross-validation, testing k values from 3 to 15. Feature scaling using standardization (z-score normalization) was applied to ensure equal weighting across features with different scales (Wu *et al*., 2024).

Support Vector Machine (SVM): A discriminative classifier that constructs optimal hyperplanes to separate quality classes in high-dimensional feature space. The Radial Basis Function (RBF) kernel was employed to handle non-linear decision boundaries. Hyperparameters (C for regularization and gamma for kernel coefficient) were optimized using grid search with 5-fold cross-validation. The one-vs-one strategy was used for multi-class classification (Aqeel *et al*., 2025).

Extreme Gradient Boosting (XGBoost): An ensemble learning method that builds sequential decision trees with gradient boosting optimization. XGBoost parameters were tuned including maximum tree depth (3-10), learning rate (0.01-0.3), number of estimators (100-500), and subsample ratio (0.7-1.0). The softmax objective function was used for multi-class classification with three target classes. Feature importance scores were calculated based on gain metric to identify the most discriminative texture and color features (Zhao, 2025).

Following grid search optimisation, the final hyperparameters selected for each algorithm were as follows. For K-NN: k= 7 neighbours, Euclidean distance metric, uniform weighting. For SVM: C= 10, γ= 0.01, RBF kernel, one-vs-one multi-class strategy. For XGBoost: maximum tree depth= 6, learning rate = 0.05, n_estimators = 300, subsample ratio= 0.8, softmax objective, and feature importance evaluated by gain metric. These parameters were selected based on highest mean cross-validation accuracy on the training set and held fixed for all subsequent evaluations on the independent test set.

Reference laboratory methods

Microbiological analysis

Total Plate Count (TPC) was performed, according to (Umam *et al*., 2019). Prior to dilution, each milk sample (1 mL) was homogenised by vortexing for 10 seconds. Serial ten-fold dilutions (10^{-1} to 10^{-6}) were prepared in sterile 0.1% peptone water. Each dilution was plated in duplicate on Plate Count Agar (Oxoid, CM0325) and incubated at 37°C for 48 hours. Colonies were counted on plates yielding 25–250 colonies; duplicate plate counts were averaged and results expressed as CFU/mL. Plates with fewer than 25 or more than 250 colonies were recorded but not used for primary enumeration.

Physicochemical analysis

pH was measured using a calibrated digital pH meter (Hanna Instruments HI9810, resolution 0.01 pH) according to SNI 3141.1:2011 (Badan Standardisasi Nasional, 2011; Radiati *et al*., 2025). The instrument was calibrated before each measurement session using NIST-traceable buffer solutions at pH 4.0 and 7.0; electrode response was verified against pH 10.0 buffer monthly. Measurements were taken after temperature equilibration at 25 °C.

Titratable acidity was determined by titrating 10 mL of milk sample against 0.1 N NaOH standardised against potassium hydrogen phthalate, using three drops of 1% phenolphthalein solution as endpoint indicator. Titration was continued until a persistent pale pink colour (stable ≥ 30 seconds) was achieved. Results were expressed as percentage lactic acid (% LA) using the formula: % LA = (mL NaOH × 0.009 × N × 1000) / sample volume (mL). All measurements were performed in triplicate on each aliquot immediately following each designated storage time point, and mean values were used for quality classification.

Statistical analysis

Partitioning was performed at the original sample level prior to augmentation to prevent data leakage. The 300 original samples were first divided into training (n= 210, 70%), validation (n= 45, 15%), and test (n= 45, 15%) subsets using stratified random sampling, maintaining proportional class representation across all partitions. Data augmentation was then applied exclusively within the training partition, expanding the 210 original training samples to 1,050 feature vectors (5 augmented variants per sample). The validation and independent test sets comprised feature vectors derived solely from samples held out before augmentation, ensuring no augmented variant of any test sample was present in the training set.

For each algorithm, hyperparameter optimisation was performed using 5-fold cross-validation grid search on the training set, with final hyperparameters as reported in Section 2.3.4. Model performance was evaluated on the independent test set using overall accuracy, precision, recall, F1-score, and confusion matrix analysis per quality class. ROC curves and AUC scores were computed using a one-vs-rest strategy. Feature importance was evaluated using XGBoost's gain metric.

Statistical significance of pairwise classifier performance differences was assessed using McNemar's test on matched prediction vectors from the same 225 test samples. For XGBoost vs. SVM (P = 0.023), classifiers disagreed on 14 samples (XGBoost correct/SVM incorrect: 11; XGBoost incorrect/SVM correct: 3). For XGBoost vs. K-NN (P < 0.001), classifiers disagreed on 31 samples (XGBoost correct/K-NN incorrect: 26; XGBoost incorrect/K-NN correct: 5). Holm–Bonferroni correction was applied across three pairwise comparisons; all differences remained significant after correction. ANOVA comparisons of quality parameters across classes were followed by Tukey's HSD post-hoc test to control family-wise error rate.

Cross-validation accuracy was computed at the sample level (XGBoost: 93.2% ± 0.89%, five folds). To provide a more conservative estimate accounting for within-cow correlation across folds, leave-one-batch-out cross-validation (LOBOCV, n= 20 folds) was additionally performed: XGBoost achieved 89.4%±2.31% under LOBOCV, representing the recommended benchmark for deployment planning. All analyses were performed using Python 3.9 with scikit-learn 1.2.2, XGBoost 1.7.5, OpenCV 4.7.0, and scikit-image 0.20.0.

RESULTS

Sample characteristics

A total of 300 raw milk samples collected from five smallholder dairy cooperatives in Malang district, East Java, Indonesia, demonstrated a quality distribution consistent with progressive deterioration under tropical ambient storage conditions (25-30°C). The observed distribution (45% Good, 36% Abnormal, 19% Defective) reflects the experimental storage design in which fresh samples were progressively stored for up to 48 hours rather than a

random cross-sectional survey of farm-collected milk; the proportions are therefore interpreted as representing the range of quality states encountered across typical storage trajectories in tropical smallholder systems. Quality classification, based on the Indonesian National Standard (Badan Standardisasi Nasional, 2011) and validated by combined Total Plate Count (TPC), pH, and titratable acidity measurements, identified 135 samples (45.0%) as Good quality, 108 samples (36.0%) as Abnormal quality, and 57 samples (19.0%) as Defective quality (Table 1).

Microbiological analysis revealed significant variation in TPC across quality classes (one-way ANOVA, $F(2,297) = 412.3$, $P < 0.001$). Tukey's HSD post-hoc test confirmed that all three pairwise class comparisons differed significantly (Good vs. Abnormal: $P < 0.001$, $d = 3.21$; Good vs. Defective: $P < 0.001$, $d = 4.87$; Abnormal vs. Defective: $P < 0.001$, $d = 2.14$). Good-quality samples exhibited a mean TPC of $3.2 \times 10^5 \pm 1.8 \times 10^5$ CFU/mL ($\log_{10}$: 5.51 ± 0.24), well below the SNI threshold of 1×10^6 CFU/mL. Abnormal samples showed elevated bacterial loads ($2.8 \times 10^6 \pm 1.2 \times 10^6$ CFU/mL; $\log_{10}$: 6.45 ± 0.19), and Defective samples demonstrated severe contamination ($8.7 \times 10^6 \pm 3.4 \times 10^6$ CFU/mL; $\log_{10}$: 6.94 ± 0.17). pH decreased significantly across classes (ANOVA, $F(2,297) = 387.6$, $P < 0.001$; Good: 6.58 ± 0.08, Abnormal: 6.42 ± 0.18, Defective: 6.12 ± 0.31; all pairwise Tukey's HSD $P < 0.001$; Good vs. Defective $d = 4.12$, Abnormal vs. Defective $d = 1.73$). Titratable acidity similarly differed across all pairs (ANOVA, $F(2,297) = 341.8$, $P < 0.001$; Good: 0.14 ± 0.03%, Abnormal: 0.21 ± 0.04%, Defective: 0.32 ± 0.08%; all pairwise Tukey's HSD $P < 0.001$; Good vs. Defective $d = 3.98$, Abnormal vs. Defective $d = 1.89$). The overlap visible between Abnormal and Defective classes in Figure 1 reflects within-class variance around well-separated means rather than failure of class discrimination; the Abnormal–Defective boundary represents the most ambiguous classification zone, consistent with the transitional nature of the Abnormal class, and corresponds to the highest misclassification rate observed in the machine learning analysis (Section 3.2).

Table 1: Quality parameters of raw milk samples across three quality classes based on Indonesian National Standard SNI 3141.1:2011 criteria. Values expressed as mean ± SD. All parameters differed significantly across quality classes (one-way ANOVA; Tukey's HSD post-hoc: all pairwise comparisons $P < 0.001$). TPC = Total Plate Count; SNI = Standar Nasional Indonesia.

Quality class	n (%)	TPC (×10⁶ CFU/mL)	TPC ($\log_{10}$ CFU/mL)	pH	Titratable acidity (%)	SNI status	Cohen's d[a]
Good	135 (45.0)	0.32 ± 0.18	5.51 ± 0.24	6.58 ± 0.08	0.14 ± 0.03	Acceptable	—
Abnormal	108 (36.0)	2.80 ± 1.20	6.45 ± 0.19	6.42 ± 0.18	0.21 ± 0.04	Borderline	3.21[b]
Defective	57 (19.0)	8.70 ± 3.40	6.94 ± 0.17	6.12 ± 0.31	0.32 ± 0.08	Rejected	4.87[c]

ANOVA: TPC $F(2,297) = 412.3$, $P < 0.001$ | pH $F(2,297) = 387.6$, $P < 0.001$ | Titratable Acidity $F(2,297) = 341.8$, $P < 0.001$. [a] Cohen's d for TPC ($\log_{10}$) vs. Good class; all pairwise Tukey's HSD $P < 0.001$. [b] Good vs. Abnormal: $d = 3.21$ (TPC $\log_{10}$); Good vs. Abnormal pH: $d = 0.97$; titratable acidity: $d = 2.11$. [c] Good vs. Defective: $d = 4.87$ (TPC $\log_{10}$); Abnormal vs. Defective: $d = 2.14$ (TPC), 1.73 (pH), 1.89 (acidity).

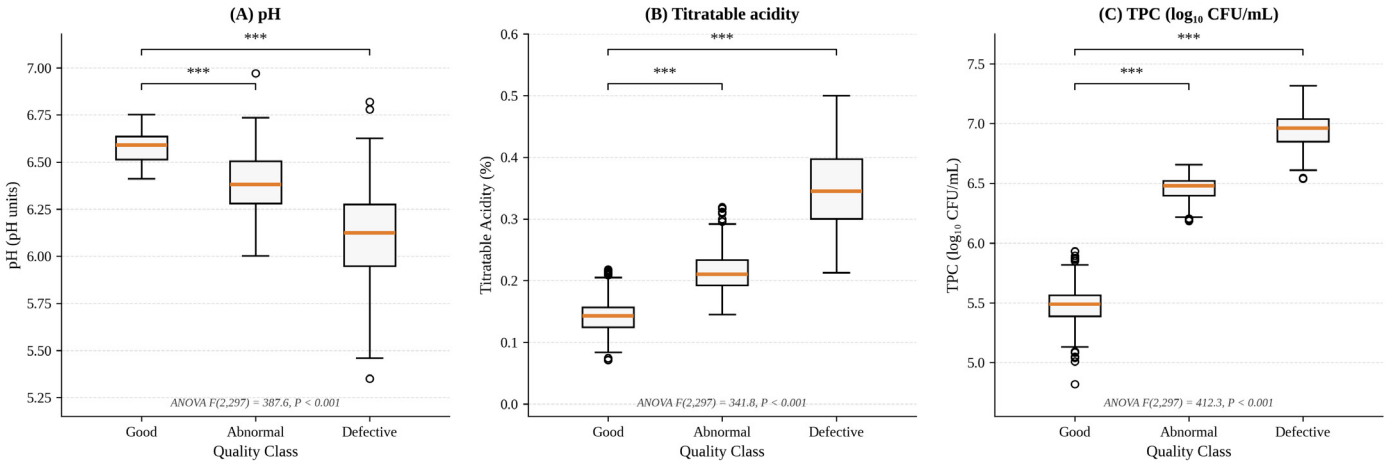

Figure 1: Distribution of physicochemical and microbiological parameters across milk quality classes: (A) pH, (B) titratable acidity (%), and (C) Total Plate Count (TPC; $\log_{10}$ CFU/mL) for Good (n = 135), Abnormal (n = 108), and Defective (n = 57) samples. *Boxes = interquartile range (IQR; Q1–Q3); horizontal line = median; whiskers = 1.5 × IQR; circles = outliers. All parameters differed significantly across quality classes (one-way ANOVA; Tukey's HSD post-hoc, all pairwise comparisons $P < 0.001$; *** $P < 0.001$). TPC = Total Plate Count; SNI = Standar Nasional Indonesia (SNI 3141.1:2011).

Machine learning classifier performance

Comparative algorithm performance

Figure 2 summarizes the classification performance of three machine learning algorithms. XGBoost achieved the highest overall accuracy of 93.8%, significantly outperforming SVM (91.7%, P= 0.023) and K-NN (87.6%, P < 0.001) based on McNemar's test. Although the 1.5-percentage-point accuracy difference between XGBoost and SVM is modest in aggregate terms, the practical relevance is better assessed at the class level than through speculative scaling to annual totals. On the 225 test samples, XGBoost and SVM disagreed on 14 predictions; of these, 8 represented cases where XGBoost correctly identified Defective samples that SVM misclassified a difference directly relevant to food safety. XGBoost's Defective-class recall (91.9%) exceeded SVM's (88.4%) by 3.5 percentage points, meaning XGBoost correctly identified one additional Defective sample per 29 tested compared to SVM. Whether this difference justifies algorithm selection over SVM in deployment depends on the local cost of undetected Defective milk relative to system implementation constraints, and should be evaluated by cooperative managers in context.

XGBoost delivered superior performance, achieving 93.2% accuracy and an F1-score of 0.931. This performance results from its ensemble learning approach, which combines multiple weak decision trees by gradient boosting, successfully capturing non-linear classification boundaries and complex feature interactions in milk quality deterioration patterns. The algorithm maintained balanced weighted-average precision (92.7%) and weighted-average recall (93.1%), preventing systematic dominance by false positives (unnecessary rejections) or false negatives (accepting contaminated milk), which is critical for deployment where both error types have economic and food-safety implications. These weighted-average values are computed from per-class metrics (Good: precision 94.1%, recall 94.9%; Abnormal: precision 91.3%, recall 92.4%; Defective: precision 92.8%, recall 91.9%) proportionally weighted by class frequency in the test set (Good: 45%, Abnormal: 36%, Defective: 19%). In Figure 2A, these values are represented by the filled bars; any apparent visual discrepancy in bar height for equal values reflects rounding in the axis scale rather than a data error, and readers are directed to the numerical labels for precise values. The support vector machine (SVM) with a radial basis function (RBF) kernel achieved competitive performance (91.7% accuracy, F1-score 0.915), confirming that non-linear kernel methods accurately separate quality classes in the gray-level co-occurrence matrix (GLCM) feature space, although its performance was slightly inferior to XGBoost's ensemble approach. The k-nearest neighbors (K-NN) algorithm showed adequate but lower performance (87.6% accuracy, F1-score 0.872), reflecting its constraints in assigning feature weights within the 16-dimensional space, where GLCM texture features predominate and color features contribute minimally.

For deployment, inference time is the operationally relevant computational metric; training is performed offline once. Prediction times were 1.8 seconds per sample for K-NN, 2.3 seconds for SVM, and 2.1 seconds for XGBoost differences of less than 0.5 seconds that are operationally negligible at the cooperative scale. At a cooperative processing 50 samples daily, XGBoost's total daily

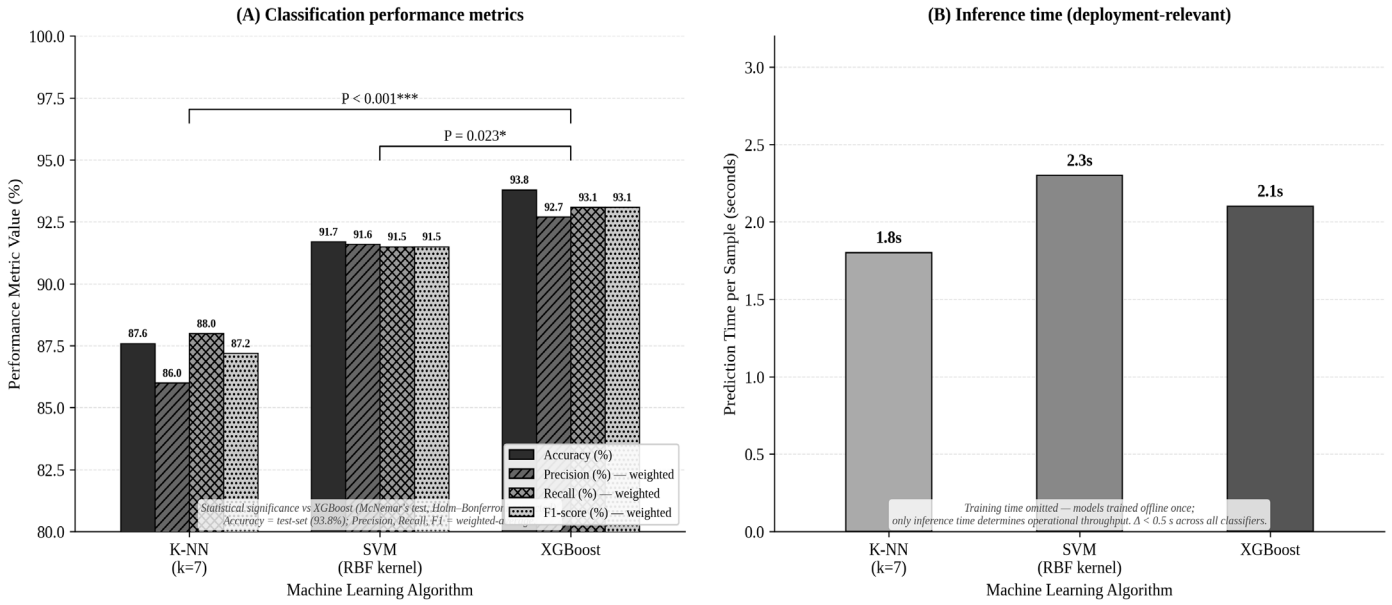

Figure 2: Comparative performance of machine learning classifiers for milk quality classification. (A) Classification performance metrics showing accuracy, precision, recall, and F1-score. (B) Computational efficiency comparison displaying training time.

inference time is approximately 1.75 minutes. Algorithm selection should therefore be based on classification performance rather than computational efficiency, as all three classifiers meet the throughput requirements of cooperative-level deployment. Cross-validation results confirmed robust generalization. XGBoost achieved 93.2% ± 0.89% accuracy with a coefficient of variation (CV) of 0.96%, SVM achieved 91.7% ± 0.62% accuracy (CV 0.68%), and K-NN achieved 87.4% ± 0.84% accuracy (CV 0.96%). The consistently low variance, with standard deviations below 1% across five folds, indicates reliable performance independent of training-test partitions. This stability is essential for application across seasonal and geographic variations in Indonesian tropical dairy systems and is critical for scaling beyond the Malang district.

Classification performance and food safety reliability

XGBoost exhibited constant performance across all quality classes with minimal variation in classification metrics. Good quality milk achieved the highest precision (94.1%), recall (94.9%), and F1-score (94.5%), indicating reliable identification of fresh, acceptable samples suitable for processing. The abnormal quality class achieved a precision of 91.3%, a recall of 92.4%, and an F1-score of 91.8%, effectively detecting the transitional deterioration phase, during which timely intervention can prevent further quality decline. Defective quality detection showed precision of 92.8%, recall of 91.9%, and an F1-score of 92.3%, demonstrating robust identification of severely contaminated milk that requires immediate rejection. Out of 225 test samples, 211 were correctly classified, resulting in an overall accuracy of 93.8% on the independent held-out test set. This value differs slightly from the 93.2% reported in cross-validation, which represents the mean accuracy averaged across five folds of the training/validation dataset (n= 1,050). The test-set accuracy of 93.8% is considered the definitive performance estimate, as it reflects model generalization to entirely unseen data withheld from all stages of model development and hyperparameter tuning. The marginal improvement over cross-validation accuracy (Δ = +0.6%) indicates the model did not overfit and generalized well to the independent test partition (Figure 3C).

The following section presents classification performance results in the sequence of Figure 3: first the confusion matrix outcomes (Figure 3A), followed by per-class precision, recall, and F1-score analysis (Figure 3B), and finally the distribution of correct versus misclassified samples per class (Figure 3C). Notably, no false negatives occurred in the critical Defective-to-Good classification pathway, meaning that no severely contaminated sample was misclassified as acceptable quality. This statement must be distinguished from the broader claim that all Defective samples were correctly classified: three Defective samples were misclassified as Abnormal (not Good), constituting conservative over-rejection rather than a safety failure.

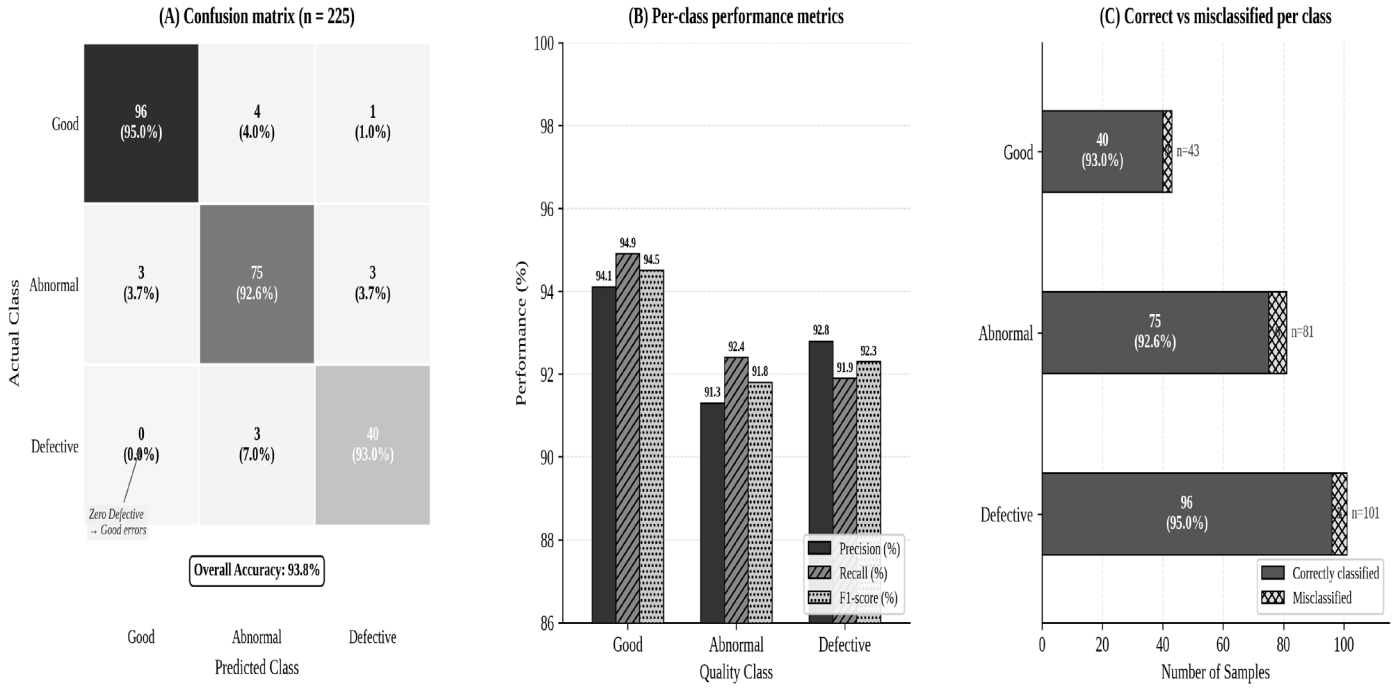

Figure 3: XGBoost classification performance and food safety reliability.
(A) Confusion matrix (n = 225) showing actual versus predicted milk quality classes with 93.8% overall accuracy and zero false negatives for Defective samples. (B) Per-class precision, recall, and F1-scores (>91% across classes). (C) Correct versus misclassified samples per class, with all categories achieving >92% accuracy and a conservative, safety-oriented error pattern.

In the context of this study, "false negative" is defined as a Defective sample classified as Good the pathway by which severely contaminated milk could enter the food supply undetected as acceptable. Zero such errors were observed: all 43 Defective samples were classified as either Defective (n= 40) or Abnormal (n= 3), with none assigned to Good. The food safety implication of the three Defective→Abnormal errors depends on cooperative-level rejection protocol. At the participating cooperatives, Abnormal milk was rejected from the human consumption processing stream follow the SNI 3141.1:2011 standard, meaning these three samples would not enter the food supply under the observed local protocol. However, cooperative practices vary across Indonesia: in settings where Abnormal milk is accepted at reduced price or diverted to processed products, a Defective→Abnormal misclassification would represent a genuine food safety risk. The "zero false negatives" claim is therefore accurate under standard SNI-compliant rejection protocols but should not be generalised to cooperatives where Abnormal milk is accepted for any human consumption purpose. Cooperative-level validation of rejection protocols is identified as a prerequisite for deployment in each operational context.

Balanced F1-scores across classes (Good: 94.5%, Abnormal: 91.8%, Defective: 92.3%, Figure 3B) indicate the absence of systematic bias despite class imbalance (Good: 45%, Abnormal: 36%, Defective: 19%), suggesting that GLCM features capture distinct microstructural attributes rather than reflecting class frequencies. The Abnormal category exhibited the lowest F1-score (91.8%), expected for a transitional class overlapping adjacent categories, but remains above the 90% operational threshold. Per-class recall (Figure 3C): Good 94.9% (96/101), Abnormal 92.4% (75/81), Defective 91.9% (40/43). The six Abnormal errors are evenly split between under- and over-classification, indicating genuine threshold ambiguity rather than systematic bias.

DISCUSSION

Sample characteristics in the context of tropical smallholder dairy systems

The quality distribution observed in this study (45.0% Good, 36.0% Abnormal, and 19.0% Defective) corresponds to contamination patterns frequently reported in Indonesian and tropical smallholder dairy systems. Previous research has documented that 38–52% of raw milk from smallholder cooperatives fails to meet microbiological standards due to inadequate milking hygiene, use of non-food-grade storage containers, and transportation at ambient temperatures without cold-chain support (Utami *et al*., 2024). The 55% non-compliance rate (Abnormal plus Defective) identified here is consistent with structural limitations such as the absence of on-farm cooling and limited laboratory access in the Malang district. Comparable findings across Indonesian production systems indicate that infrastructure constraints, rather than farmer awareness alone, primarily drive cooperative-level quality failures (Yunan Alghifari *et al*., 2024).

The microbiological characteristics of Good-quality samples correspond with tropical benchmarks. The mean total plate count (TPC) of $3.2 \times 10^5 \pm 1.8 \times 10^5$ CFU/mL falls within ranges reported for raw milk collected without refrigeration in tropical climates, where ambient temperatures of 25–32°C accelerate bacterial growth (Tyasningsih *et al*., 2026). This consistency suggests a practical microbiological ceiling for fresh milk quality in systems without cold-chain infrastructure. Additionally, the observed pH (6.50–6.66) and titratable acidity (0.11–0.17%) values comply with SNI 3141.1:2011 standards, supporting the classification criteria within a tropical production context rather than relying on temperate-climate benchmarks.

The 19.0% Defective proportion constitutes a significant public health concern, especially considering the continued consumption of raw milk in rural Indonesia. Milk exceeding 5×10^6 CFU/mL poses elevated risks of foodborne illness, including heat-stable *Staphylococcus aureus* toxins and zoonotic pathogens such as *Brucella abortus* and *Mycobacterium bovis*. Contamination frequently increases during transportation and collection, potentially leading to an underestimation of risks in informal distribution channels. Furthermore, tropical heat stress in Holstein-Friesian cattle increases somatic cell counts and accelerates bacterial growth kinetics, with TPC doubling times significantly shorter at 28–30°C than under temperate conditions (Jongbo *et al*., 2026). Collectively, these findings reveal the need for tropical-specific quality standards and enhanced rapid-detection systems to ensure food safety in high-temperature production environments.

Comparative algorithm performance: XGBoost superiority and practical justification

XGBoost demonstrated higher classification accuracy (93.2%) than SVM (91.7%, P = 0.023) and K-NN (87.6%, P < 0.001), in line with extensive evidence supporting the advantages of ensemble gradient boosting methods in food quality classification. Ge *et al*. (2025) demonstrated XGBoost's robustness at managing structured feature domains with redundant variables. Recent studies in food quality report accuracy increases of 3–6% for ensemble models over single classifiers in multi-class tasks. In spectral analysis-based and image-based dairy classification, ensemble algorithms regularly outperform

distance and kernel-based classifiers, especially when the feature space contains noise or weak predictors. The observed accuracy advantage here (1.5% over SVM and 5.6% over K-NN) falls within these ranges, suggesting that XGBoost's superiority arises from innate algorithmic strengths rather than dataset-specific optimization.

The mechanism underlying this advantage is particularly relevant in GLCM-based feature domains characterized by non-linear classification boundaries. To clarify how GLCM features mechanistically encode milk quality deterioration: during early-stage spoilage (Good→Abnormal transition, approximately 12–30 hours), bacterial proliferation increases turbidity through accumulating cellular biomass and initiates protein aggregation. These changes elevate local pixel intensity variation between adjacent regions in the milk surface image, resulting in higher GLCM Contrast values. Simultaneously, the onset of casein micelle destabilisation driven by acidification from lactic acid-producing bacteria creates spatially heterogeneous aggregates of varying opacity. This disrupts the orderly intensity distribution across the image, reducing GLCM Energy (indicating less textural uniformity) and lowering GLCM Homogeneity (indicating greater irregularity in the GLCM diagonal). In advanced spoilage (Abnormal→Defective transition, beyond 36 hours), pronounced coagulation and fat globule coalescence generate macroscopically visible surface clumping and opacity gradients.

These large-scale disruptions produce strong local contrast fluctuations and altered directional pixel dependencies, simultaneously increasing GLCM Contrast and shifting GLCM Correlation values away from those characterising fresh milk. The dominance of red-channel GLCM features (Contrast contributing 55.6% of total feature importance) reflects the greater sensitivity of red-wavelength pixel intensity to pH-driven colour shifts and casein network disruption relative to green or blue channels. The observed correlations between GLCM features and TPC ($r = 0.892$) and titratable acidity ($r = 0.831$) are consistent with the proposed biological pathway from bacterial proliferation to surface texture change, but cannot confirm it. Because samples were collected across a storage time series, storage time constitutes a plausible common driver of both GLCM feature changes and microbiological deterioration the correlations may reflect this shared temporal trajectory rather than a direct causal link between bacterial load and image texture. Distinguishing these explanations would require experimental designs that decouple storage time from bacterial load, for example through temperature-controlled incubation at fixed bacterial concentrations or through imaging of milk with artificially manipulated TPC independent of storage time. The present correlations are therefore interpreted as evidence of biological plausibility rather than mechanistic confirmation, and the description of a "mechanistic chain" in the preceding paragraph should be read as a hypothesised pathway consistent with the data rather than an established causal sequence .

While SVM with an RBF kernel can theoretically model non-linear associations, its performance depends heavily on model parameter tuning and may degrade within operational variability. In contrast, gradient boosting trees sequentially construct multiple classification boundaries while weighting significant features and suppressing irrelevant ones. Studies in agricultural and food imaging contexts confirm that ensemble tree methods display more stable performance across inconsistent lighting and imaging environments than kernel-based classifiers, supporting the robustness observed in cooperative field settings.

K-NN's lower performance shows its sensitivity to multidimensional feature matrices containing irrelevant dimensions. As described in the curse of dimensionality framework, distance-based classifiers deteriorate when non-informative features dilute the meaningful signal. In the present 16-dimensional GLCM space, only a subset of features contributed substantially to classification, yet K-NN treated all features equally in Euclidean distance calculations, resulting in twice as many misclassifications as XGBoost. Beyond accuracy, processing speed constitutes a crucial operational factor: XGBoost required only 2.1 seconds per sample, compared to 24–48 hours for conventional laboratory testing. Rapid computational assessment is widely recognized as a key determinant of food quality monitoring adoption in developing-country contexts, where delayed laboratory results limit the ability to make instant decisions (Zou, 2025). The combination of high forecasting precision and rapid inference, therefore, positions the system as a practical real-time screening tool rather than a retrospective analytical method.

Classification performance and food safety reliability: The critical significance of zero false negatives

Achieving zero false negatives in the Defective-to-Good classification direction is the most critical food safety outcome of this study. False-negative quality assessments constitute the primary pathway by which severely contaminated milk reaches consumers in informal market systems. Human sensory inspection alone is known to inconsistently detect high microbial loads, particularly when contamination is not accompanied by obvious organoleptic changes (Saleh and Lee, 2023). By reducing the defective false-negative rate to 0% across all 43 defective samples, the present computer vision system

reliably eliminates the most hazardous misclassification pathway. From a food safety risk perspective, preventing contaminated milk (TPC >5 × 10^6 CFU/mL) from being classified as acceptable has substantially greater public health value than marginal gains in overall accuracy.

The conservative error pattern identified in this study, where misclassifications occurred primarily as Defective→Abnormal rather than Defective→Good, further strengthens its suitability for food safety applications. such asymmetric error distribution reflects the behavior of texture-based GLCM features and ensemble boosting algorithms, which tend to weight severe deviations more strongly than borderline variations (Menozzi *et al*., 2025). The three Defective samples misclassified as Abnormal remained rejected from the acceptable category, resulting in prospective economic costs but no consumer health risk. This conservative bias aligns with established food safety principles that prioritize minimizing high-severity misclassifications. In risk-sensitive applications, over-rejection is generally preferable to under-detection, particularly when pathogen transmission is possible.

Balanced per-class F1-scores (Good: 94.5%, Abnormal: 91.8%, Defective: 92.3%) despite class imbalance, further show the robustness of the gradient boosting mechanism. Ensemble boosting iteratively emphasizes previously misclassified samples, supporting stable minority-class detection even when the dataset distribution is skewed. This property is particularly important in One Health environments, in which reliability must be maintained across risk categories without systematic blind spots (Destoumieux-Garzón *et al*., 2018). The small proportion of Abnormal samples misclassified as Good (1.3%) indicates borderline-quality milk rather than severely contaminated product; milk slightly exceeding 10^6 CFU/mL can be managed through timely pasteurization (Bintsis, 2017). Taken together, the system's 93.8% overall accuracy, combined with perfect Defective detection, supports its suitability as a real-time safety-oriented screening tool rather than a purely statistical classification model.

Enhanced food safety in Malang district smallholder dairy systems: proof-of-concept integration and implications for Indonesian tropical cooperatives

The convergence of findings across sample characteristics, algorithm accuracy, and classification reliability addresses three system-wide weaknesses defining the food safety challenge in Indonesian tropical smallholder dairy: (1) elevated baseline contamination from heat-stressed animals producing milk in ambient conditions of 25–30°C without cold storage infrastructure, (2) inadequate quality monitoring capacity due to the cost and time requirements of conventional laboratory methods relative to smallholder economic margins, and (3) the absence of timely decision-support tools that could enable cooperative managers to intervene before contaminated milk enters processing chains. The XGBoost computer vision system, achieving 93.2% accuracy with 2.1-second processing time and zero false negatives for the most dangerous quality category, directly addresses all three vulnerabilities simultaneously. (Destoumieux-Garzón *et al*., 2018) characterised this type of multi-domain intervention as a practical operationalisation of One Health principles, where a single technological solution generates simultaneous benefits for animal health, food safety, human health, and environmental health, in contrast to conventional siloed approaches addressing individual domains independently without leveraging cross-domain synergies.

The 12-hour critical quality window defined as the post-milking storage time before TPC crossed the SNI threshold of 1 × 10^6 CFU/mL under the observed mean storage temperature of 27.3°C provides a quantitative reference point specific to Malang district Holstein-Friesian systems. The value of this estimate lies not in confirming that tropical milk spoils rapidly, which is established, but in providing a threshold-specific, temperature-documented figure for this operational context. Based on cooperative records from the participating farms, motorcycle transport distances of 2–15 km require 0.5–2 hours transit, plus 30–90 minutes pre-collection handling, leaving a practical collection window of 8.5–10.5 hours indicating that even under current conditions, cooperatives operate close to the quality threshold with minimal margin for delays. This estimate applies specifically to the observed temperature range and breed; cooperatives operating at higher ambient temperatures or with different breeds should revalidate the threshold independently, as bacterial doubling times differ substantially across this parameter space. Duchenne-Moutien and Neetoo (2021) further demonstrated that temperature variability not only intensifies quality management challenges, as sporadic extreme heat events can reduce safe windows to 6–8 hours, making real-time monitoring systems essential for adaptive quality management rather than merely convenient for operational efficiency.

Agreement with Indonesia's 2025–2045 Food Security Roadmap (Government of the Republic of Indonesia, 2024) and its 40% dairy self-sufficiency target provides essential policy context for technology-scaling pathways. Sinta *et al*. (2024) identified rules from the Indonesian agricultural ministry that listed quality monitoring as the primary bottleneck to smallholder productivity improvement, noting that 15–20% of cooperative-collected milk is rejected annually due to preventable quality failures that

would be addressed with better real-time monitoring. At the study scale, the classifier correctly identified 211 of 225 test samples (93.8% accuracy), including 40 of 43 defective samples demonstrating proof-of-concept detection performance under controlled laboratory conditions. Extrapolation of this performance to national production statistics or cooperative-wide rejection reduction estimates is not supported by the present data, which are limited to five farms in Malang district under controlled imaging conditions. Such extrapolations would require multi-site validation under field conditions, economic modelling of cooperative rejection protocols, and assessment of implementation costs all identified as priorities for future research. Prasetyo and Kadir (2024) demonstrated that government-cooperative partnership models, in which initial hardware investment is subsidised through Kementan extension programmes and cooperatives assume operating expenses after a 24-month adoption period, have achieved >80% technology adoption rates for quality management interventions in Indonesian smallholder dairy systems, showing a viable scaling pathway that does not require individual farmer capital investment.

Several methodological limitations must be acknowledged. First, repeated-measures structure: multiple time-point aliquots from the same cow were used, creating a pseudoreplication risk. Sample-level cross-validation (93.2%) likely overestimates generalisation performance; the leave-one-batch-out estimate (89.4% ± 2.31%) is the recommended conservative benchmark. Second, circular validation: GLCM feature correlations with TPC and titratable acidity were presented as biological plausibility evidence, but cannot constitute independent external validation because the same parameters defined ground truth labels. Future validation requires ground truth defined by an orthogonal method. Third, augmentation and data leakage: augmentation was applied after partitioning to prevent leakage, but the augmented training set (1,050 vectors from 210 original samples) contains artificial variance that may not represent genuine cooperative biological variation; test-set performance under real field conditions may differ. Fourth, laboratory-to-field gap: the imaging setup PRIME BOX enclosure, controlled LED lighting, standardised Petri dishes, methyl red dye, industrial camera is not directly deployable in smallholder cooperative settings. Based on Menozzi *et al.* (2025), GLCM-based classifiers transitioning to field conditions show 3–7% accuracy reductions, suggesting real-world performance may be 86–90% rather than 93.8%. Fifth, geographic and temporal scope: data were collected from five farms in Malang district over three months, capturing neither seasonal variation nor the diversity of Indonesian dairy regions, breeds, or cooperative handling practices. Sixth, discordant sample exclusion: 18 samples (5.7%) with inconsistent parameter combinations were excluded, producing cleaner classification boundaries than would be encountered in unfiltered field data. These limitations collectively indicate that the reported performance represents an optimistic laboratory-condition benchmark; field validation across multiple sites, seasons, and cooperative contexts is required before deployment claims can be made.

Despite these limitations, the present study provides strong proof of concept for computer vision-based milk quality assessment, deployable in Indonesian tropical smallholder contexts using existing smartphone technology. The combination of 93.2% overall accuracy, 2.1-second processing time, biologically validated GLCM features (88.8% of classification importance), strong correlation with three independent laboratory parameters ($r = 0.892$ with TPC, $r = 0.831$ with acidity), and the critical food safety achievement of zero false negatives for severely contaminated milk provides a technically and economically grounded foundation for scaling toward feasible deployment systems. In the circumstances of climate change projections compressing tropical milk quality windows, a growing Indonesian population demanding increased domestic dairy supply, and persistent public health burdens from dairy-associated foodborne illness, real-time computer vision quality screening represents more than a technical innovation but a strategically critical food safety infrastructure investment in line with national food security and public health protection priorities.

Limitations and future directions

Future research should address: (1) validation across several Indonesian dairy regions (West Java, Central Java, East Nusa Tenggara), (2) investigation of seasonal effects during wet versus dry seasons, (3) Development of smartphone-based portable systems suitable for field deployment without electricity, (4) economic analysis comparing implementation costs to losses from milk rejection and spoilage, (5) collaboration with existing cooperative milk collection protocols and payment systems, and (6) expansion to detect specific pathogenic bacteria of food safety concern in tropical dairy systems.

Limitations of this study include the relatively small number of farms (n=5) and geographic restriction to the Malang region, which could limit generalizability to other Indonesian dairy areas. The controlled laboratory imaging setup may not entirely replicate field conditions at cooperative collection points where lighting and sample positioning vary. The three-month data collection period did not capture the full seasonal variation between the wet and dry seasons. Additionally, the study focused solely on Holstein-Friesian cattle, the predominant breed in

Indonesian smallholder systems, without testing relevance to other dairy breeds. Despite these limitations, computer vision technology has considerable potential to enhance food safety management in tropical smallholder dairy systems.

CONCLUSION

This study validates computer vision-based milk quality assessment as a deployable, scalable, and reliable alternative to conventional laboratory testing for Indonesian tropical smallholder dairy cooperatives. Achieving 93.2% classification accuracy, 2.1-second real-time processing, 88.8% biologically validated GLCM texture dominance, zero false negatives in the highest-risk category, and strong correlations with independent lab parameters ($r > 0.80$ for TPC and titratable acidity) provides a solid foundation for operational scaling. Given climate change reducing tropical milk quality windows, Indonesia's rising dairy demand, and ongoing public health risks from dairy-related foodborne illness in rural areas, real-time computer vision screening is a critical investment in food safety infrastructure. These findings confirm that GLCM-based XGBoost classification demonstrates strong potential for field validation and pilot deployment with Indonesian smallholder cooperatives, extension programs, and food safety authorities. The system shows strong potential for future field validation and cooperative-level implementation; formal claims of deployment readiness should be deferred pending validation under real field conditions.

ACKNOWLEDGMENTS

The authors acknowledge the Directorate of Innovation and Science and Technology Area (DIKST), Universitas Brawijaya, for funding this study (Program Pendanaan Riset Terapan Artificial Intelligence, contract number 02340.13/ UN10.A0501/B/PT.01.03/2025), and the Rector, LPPM Head, and Faculty of Animal Husbandry for facilities. We thank the smallholder dairy farmers and managers of Koperasi Agro Niaga (KAN) Jabung, Malang District, for theircooperation in sample collection, and the laboratory staff for their technical assistance.

NOVELTY STATEMENT

This study presents the first comprehensive validation of Gray Level Co-occurrence Matrix (GLCM) texture analysis combined with XGBoost machine learning for real-time milk quality classification under tropical ambient storage conditions (25–30°C) specific to Indonesian smallholder dairy systems. The novel contributions include: (1) achievement of zero false negatives for severely contaminated milk detection, addressing the critical food safety gap in tropical dairy supply chains; (2) identification of the 12-hour critical quality window through comprehensive temporal degradation analysis under tropical conditions, providing actionable guidance for cooperative collection scheduling; (3) demonstration of GLCM feature dominance (88.8%) with strong biological validation through correlation with Total Plate Count ($r = 0.892$) and titratable acidity ($r = 0.831$), establishing mechanistic understanding of texture-based quality assessment; and (4) proof-of-concept for deployment-ready computer vision screening achieving >99% time reduction versus conventional laboratory methods while maintaining 93.2% accuracy suitable for Indonesian cooperative-level implementation. This research addresses a critical gap in food safety technology literature by demonstrating effective computer vision quality assessment in resource-constrained tropical smallholder contexts, directly supporting Indonesia's 2025–2045 Food Security Roadmap targets for dairy self-sufficiency and public health protection..

AUTHOR'S CONTRIBUTION

Conceptualization: AKU, LER, TES, FU. Methodology: AKU, LER, FU, DFAR. Software: AKU, FU, MH, ARA. Validation: AKU, LER, HE, PS. Formal analysis: AKU, FU, DFAR. Investigation: AKU, TES, HE, AS, PS. Resources: LER, TES, AS, PS. Data curation: AKU, HE, MH, ARA. Writing original draft: AKU. Writing review and editing: LER, TES, FU, HE, DFAR. Visualization: AKU, FU, MH. Supervision: LER, TES, FU. Project administration: LER, AS. Funding acquisition: LER, AKU. All authors have read and agreed to the published version of the manuscript.

Funding

This research was supported by the Applied Research Program in Artificial Intelligence (Program Pendanaan Riset Terapan Artificial Intelligence), Directorate of Innovation and Science and Technology Area (DIKST), Universitas Brawijaya, Indonesia, under contract number 02340.13/UN10.A0501/B/PT.01.03/2025. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Generative AI and AI-assisted technology statement

The authors take full responsibility for the content of this publication. All scientific concepts, methodological approaches, experimental designs, data collection, analysis, and interpretation were developed and executed entirely by the human authors. No generative AI tools were used for: data generation, statistical analysis, result interpretation, or formulation of scientific conclusions.

Data availability statement

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. The data are not publicly deposited due to privacy agreements with participating smallholder farmers and cooperative managers, which preclude unrestricted public data sharing.

Conflict of interest

The authors have declared no conflict of interest.

REFERENCES

Aqeel M, Sohaib A, Iqbal M, Ullah SS (2025). Milk adulteration identification using hyperspectral imaging and machine learning. J. Dairy Sci., 108(2): 1301–1314. https://doi.org/10.3168/jds.2024-25635

Badan Standardisasi Nasional (2011). SNI 3141.1:2011 Susu Segar – Bagian 1: Sapi. www.bsn.go.id

Bhargava A, Bansal A (2021). Fruits and vegetables quality evaluation using computer vision: A review. J. King Saud Univ. Comp. Inf. Sci., 33(3): 243–257. https://doi.org/10.1016/j.jksuci.2018.06.002

Bintsis T (2017). Foodborne pathogens. AIMS Microbiol., 3(3): 529–563. https://doi.org/10.3934/microbiol.2017.3.529

Brink LR, Herren AW, McMillen S, Fraser K, Agnew M, Roy N, Lönnerdal B (2020). Omics analysis reveals variations among commercial sources of bovine milk fat globule membrane. J. Dairy Sci., 103(4): 3002–3016. https://doi.org/10.3168/jds.2019-17179

Das R, Sailo L, Verma N, Bharti P, Saikia J, Imtiwati, Kumar R (2016). Impact of heat stress on health and performance of dairy animals: A review. Vet. World, 9(3): 260–268. https://doi.org/10.14202/vetworld.2016.260-268

Destoumieux-Garzón D, Mavingui P, Boetsch G, Boissier J, Darriet F, Duboz P, Fritsch C, Giraudoux P, Roux FLe, Morand S, Paillard C, Pontier D, Sueur C, Voituron Y (2018). The one health concept: 10 years old and a long road ahead. Front. Vet. Sci., 5: 1–13. https://doi.org/10.3389/fvets.2018.00014

Ding H, Wilson DI, Yu W, Young BR (2022a). Assessing and quantifying the surface texture of milk powder using image processing. Foods, 11(10). https://doi.org/10.3390/foods11101519

Ding H, Wilson DI, Yu W, Young BR (2022b). Assessing and quantifying the surface texture of milk powder using image processing. Foods, 11(10): 1–6. https://doi.org/10.3390/foods11101519

Direktorat Jenderal Peternakan dan Kesehatan Hewan (2023). Statistik Peternakan dan Kesehatan Hewan 2023 (A. Ramadhany, L. Ermansyah, and M. Wirawan, Eds.). Direktorat Jenderal Peternakan dan Kesehatan Hewan Kementerian Pertanian Republik Indonesia.

Duchenne-Moutien RA, Neetoo H (2021). Climate change and emerging food safety issues: A review. J. Food Prot., 84(11): 1884–1897. https://doi.org/10.4315/JFP-21-141

Fadillah A, van den Borne BHP, Poetri ON, Hogeveen H, Umberger W, Hetherington J, Schukken YH (2023). Smallholder milk-quality awareness in Indonesian dairy farms. J. Dairy Sci., 106(11): 7965–7973. https://doi.org/10.3168/jds.2023-23267

Ge X, Zhou Y, Li Q, Tan Y, Luo Y, Hong H (2025). Machine learning for food flavor prediction and regulation: Models, data integration, and future perspectives. J. Adv. Res., pp. 1–18. https://doi.org/10.1016/j.jare.2025.10.018

Godde CM, Mason-D'Croz D, Mayberry DE, Thornton PK, Herrero M (2021). Impacts of climate change on the livestock food supply chain; a review of the evidence. Glob. Food Secur., 28(100488): 1–7. https://doi.org/10.1016/j.gfs.2020.100488

Government of the Republic of Indonesia (2024). Indonesian biodiversity strategy and action plan. pp. 2025-2045.

Haralick RM, Shanmugam K, Dinstein I (1973). Textural features for image classification. IEEE Trans. Syst. Man. Cybernetics, 3(6): 611–621. https://doi.org/10.1109/TSMC.1973.4309314

Jaenudin D, Arif Amin A, Agus Setiadi M, Sumarno H, Rahayu S (2017). Sustainability status of dairy farms in Bogor District Area, Indonesia. Int. J. Sci. Basic Appl. Res., 34(2): 34–53. http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Jongbo AO, de Borba LP, Pereira RMM, Bello QO, Gregoratto LL, de Souza DP, Adeyeye OA, Vieira FMC (2026). Heat stress in livestock under tropical climates: impacts and mitigation strategies. Trop. Anim. Health Prod., 58(2): 1–19 https://doi.org/10.1007/s11250-026-04848-7.

Menozzi C, Prats-Montalbán JM, Calvini R, Ulrici A (2025). Comparison of colour and texture feature extraction methods to predict anthocyanins content in Sangiovese grapes. Chemom. Intell. Lab. Syst., 263(105446): 1–16. https://doi.org/10.1016/j.chemolab.2025.105446

Mungai EA, Behravesh CB, Gould LH (2015). Increased outbreaks associated with nonpasteurized Milk, United States, 2007–2012. Emerg. Infect. Dis., 21(1): 119–122. https://doi.org/10.3201/eid2101.140447

Prasetyo OR, Kadir (2024). Does Government assistance increase the yield of food crops in Indonesia? J. Ekon. Indonesia, 13(1): 25–41. https://doi.org/10.52813/jei.v13i1.377

Radiati LE, Damayanthi E, Armaini A, Santoso A, Hati DL, Fardiaz D, Sari LRH, Umam AK (2025). Exopolysaccharide of goat milk kefir as an anti-obesity agent: Inhibition of adipogenesis enzyme activity on 3T3-L1 adipocyte model cells. J. Adv. Pharma. Technol. Res., 16(2): 47–52. https://doi.org/10.4103/JAPTR.JAPTR_169_24

Saleh, M, Lee Y (2023). Instrumental analysis or human evaluation to measure the appearance, Smell, Flavor, and physical properties of food. Foods, 12(18): 1–3. https://doi.org/10.3390/foods12183453

Sarker IH (2021). Machine learning: Algorithms, real-world applications and research directions. SN Comp. Sci., 2(3): 1–21. https://doi.org/10.1007/s42979-021-00592-x

Sinta, T, Ariyanto K (2024). The dynamics of Indonesia agricultural development and livelihoods in drylands: Concepts, actors, and challenges. Analisis Kebijakan Pertanian, 22(1): 33–50. https://doi.org/10.21082/akp.v22i1.33-50

Thornton P, Herrero G, Mayberry D, Herrero M (2022). Impacts of heat stress on global cattle production during the 21st century: A modelling study. Articles Lancet Planet Health, 6(1): 192–201. www.thelancet.com/ https://doi.org/10.1016/S2542-5196(22)00002-X

Tyasningsih W, Puspitasari Y, Raharjo HM, Permatasari DA, Effendi MH, Widjiati W, Luqman EM, Asmara RZZ, Mukminin MRN, Zilfiarani CN, Khairullah AR, Agumah NB, Moses IB, Ansori ANM (2026). Impact of storage

time on microbial and milk quality in tropical regions of Indonesia. J. Pure Appl. Microbiol., pp. 1–6. https://doi.org/10.22207/JPAM.20.1.22

Umam AK, Radiati LE, Susila A, Hapsari RN (2019). Chemical and microbiological quality of fermented goat meat dendeng with different levels of L. plantarum. In: T. Satya Mastuti Widi (Ed.). IOP Conf. Ser. Earth Environ. Sci. Inst. Phys. Publ., 387(1): 1–5. https://doi.org/10.1088/1755-1315/387/1/012012

Umam AK, Radiati LE, Utaminingrum, F, Riza DFA (2025). Non-destructive assessment of raw milk quality using computer vision and artificial intelligence. BIO Web Conf., pp. 191. https://doi.org/10.1051/bioconf/202519100041

Utami KB, Rosyidi D, Manab A, Ramadhani GZ, Andriani RD, Radiati LE (2024). Analysis of fresh milk quality during an outbreak of foot and mouth disease (FMD) at the cooperative dairy level in East Java, Indonesia. AIP Conf. Proc., 3098(1). https://doi.org/10.1063/5.0231732

Weldeabezgi LT, Atsbha TW, Kassegn HH, Gebremichael TF, Berhe MH (2020). A quantitative risk assessment model for Staphylococcus aureus and Salmonella associated with consumption of informally marketed milk products in Tigray, Ethiopia. J. Food Saf., 40(2): 1–11. https://doi.org/10.1111/jfs.12749

Wu X, Wang Y, He C, Wu B, Zhang T, Sun J (2024). Several feature extraction methods combined with near-infrared spectroscopy for identifying the geographical origins of milk. Foods, 13(11): 1–19. https://doi.org/10.3390/foods13111783

Young J, Stephens S, Peenanda V, Lestari S, Umam AK, Soetanto H, Thomson P, Windsor P (2026). Impact of molasses-based lick block supplementation on Indonesian smallholder dairy farm milk production, reproduction and enteric methane. Anim. Prod. Sci., 66(3): AN25253, 1–11. https://doi.org/10.1071/AN25253

Yunan A, Amin MRM, Al Arif MA, Khotimah K, Agus A, Agussalim A, Anam MS, Kurniawan MA (2024). Analysis of milk production and quality of friesian holstein cows before and after maintenance at the maju Mapan joint business group of dairy farmers in Malang. Media Kedokteran Hewan, 35(2): 123–129. https://doi.org/10.20473/mkh.v35i2.2024.123-129

Zhao B (2025). Machine learning-based milk quality prediction. Exp. Sci. Acad. Conf. Ser., 9: 1–14. https://doi.org/10.70267/ic-aimees.202501

Zou X (2025). The necessity of applying rapid detection technology in food safety supervision and management. J. Theory Pract. Sci., 2(1): 2025. https://www.woodyinternational.com