ADBA Computer Science
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8544
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.63-71
https://doi.org/10.69882/adba.cs.2026019

Predictive Modeling for Milk Quality Using Machine Learning and XAI Algorithms

Zeynep Çetinkaya [α,1], **Fahrettin Horasan** [α,2], **Hüseyin Aydilek** [β,3] **and Mustafa Yasin Erten** [β,4]
[α]Department of Computer Engineering, Kırıkkale University, Kırıkkale, Türkiye, [β]Department of Electrical & Electronics Engineering, Kırıkkale University, Kırıkkale, Türkiye.

ABSTRACT Milk quality assessment is of critical importance for food safety and public health. Traditional milk quality evaluation relies on physicochemical measurements that require expert interpretation and may not directly support rapid or large-scale decision-making, increasing the need for data-driven and automated assessment methods. Although machine learning-based approaches have been widely applied in milk quality classification in recent years, the lack of transparency in model decision mechanisms and insufficient reporting of data preprocessing and data leakage control procedures pose significant limitations in terms of reliability. In this study, the milk quality classification performance of various supervised machine learning algorithms is comparatively evaluated using an open-access milk dataset. Random Forest (RF), k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Extreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), and Artificial Neural Network (ANN) models are assessed under fair and consistent experimental conditions. The main contribution of this study lies in the application of group-based data partitioning strategies to prevent data leakage, rather than directly removing duplicate or highly similar records from the dataset. This approach prevents data loss and enables a more realistic evaluation of model performance. Furthermore, a targeted and minimalist preprocessing strategy is adopted by applying scaling exclusively to continuous variables. For hyperparameter optimization, Grid Search and Particle Swarm Optimization (PSO) methods are employed; notably, tree-based models optimized using PSO demonstrate more consistent classification performance. To move beyond predictive accuracy, Explainable Artificial Intelligence (XAI) approaches are utilized to enhance the interpretability of model decisions. In this context, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) methods are applied to analyze the contributions of key features influencing milk quality. Experimental results indicate that physicochemical properties, particularly pH, fat content, and temperature, play a decisive role in milk quality prediction. In conclusion, this study demonstrates that the combined use of machine learning and explainability techniques provides significant contributions in terms of reliability, transparency, and methodological robustness in milk quality classification.

KEYWORDS
XAI
Optimization algorithms
Machine learning
Deep learning
Milk quality

INTRODUCTION

The food industry holds strategic importance on a global scale, both in terms of economic growth and public health. The safety, nutritional value, and sensory acceptability of food products offered on the market directly affect not only consumer satisfaction but also public health and sustainable development. In this context, food quality emerges as a decisive factor in shaping responsible production and consumption practices and plays a critical role in protecting social welfare (Peri 2006).

Within this framework, milk is regarded as a basic and strategic food product due to its high nutritional value and widespread consumption worldwide (Polat *et al.* 2021). However, milk's susceptibility to microbial spoilage and adulteration necessitates the effective and reliable implementation of quality control processes (Azad and Ahmed 2016; Polat *et al.* 2021). Therefore, milk quality assessment is considered a critical step in both safeguarding consumer health and ensuring the reliability of the food supply chain (Murphy *et al.* 2016). Traditional milk quality control methods rely on manual inspections, physical tests, and chemical analyses, and often involve time-consuming, costly, and error-prone procedures (Azad and Ahmed 2016). This situation creates significant constraints in meeting the modern food industry's increasing demands for speed, accuracy, and sustainability.

Recent advances in data science and machine learning have created significant opportunities for the automated, rapid, and reliable assessment of milk quality (Sarveswaran *et al.* 2023; Sunithamani *et al.* 2024). Studies in the literature demonstrate that machine learning-based approaches can determine milk quality with high accuracy while reducing operational costs (Sarveswaran *et al.* 2023; Manisha and Jagadeeshwar 2023). These methods enhance the effectiveness of quality control processes by minimizing human-induced errors. However, most of these high-performance models operate as "black boxes" and fail to provide sufficient explanations of their underlying decision-making mechanisms (Ribeiro *et al.* 2016).

In application areas that directly affect public health, such as food safety, high predictive accuracy alone is insufficient. It is also essential that model decisions are transparent, interpretable, and justifiable (Arrighi *et al.* 2025). Otherwise, the opacity of decision-making processes may lead to trust issues and hinder the adoption of AI-supported systems in real-world applications.

Manuscript received: 2 December 2025,
Revised: 22 January 2026,
Accepted: 25 January 2026.

[1]zeynepcetinkaya@kku.edu.tr (**Corresponding author**)
[2]fhorasan@kku.edu.tr
[3]huseyinaydilek@kku.edu.tr
[4]mustafaerten@kku.edu.tr

In this context, Explainable Artificial Intelligence (XAI) approaches offer an important solution by rendering the decision-making processes of machine learning models more transparent. XAI methods enable the analysis of which variables influence model outputs and to what extent, thereby increasing user trust and ensuring decision verifiability (Ribeiro *et al.* 2016). The integration of XAI into food quality assessment systems allows domain experts to interpret model results more reliably and enhances the practical applicability of these systems (Arrighi *et al.* 2025; Chowdhury *et al.* 2024).

In this study, a comprehensive analysis of milk quality classification was conducted using multiple machine learning algorithms, based on fundamental milk quality parameters reported in the literature. Random Forest (RF), k-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Extreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), and Artificial Neural Network (ANN) models were evaluated comparatively. Furthermore, the interpretability of model decisions was enhanced using XAI techniques, namely LIME and SHAP, through which critical variables influencing milk quality were examined in detail. In particular, SHAP analyses revealed the contribution of each feature to the model predictions and their relative importance in the decision-making process. The findings of this study are expected to contribute to improving food safety in the dairy industry and to supporting artificial intelligence-driven digital transformation processes. The remainder of this article is organized as follows. Section II reviews related studies on milk quality classification reported in the literature. Section III describes the dataset, data preprocessing procedures, and machine learning methods employed. Section IV presents the experimental results and comparative model performance analyses. Section V provides a comparative discussion of the proposed approach with existing studies. Finally, the concluding section summarizes the main findings and offers recommendations for future research.

LITERATURE REVIEW

The rapid, objective, and reproducible assessment of milk quality is of critical importance for food safety, consumer health, and quality assurance in the dairy industry. Traditional chemical and microbiological analyses are often time-consuming, costly, and highly dependent on operator expertise, which limits their effectiveness in large-scale and real-time quality control processes. Consequently, a growing body of research has emerged in recent years focusing on machine learning, deep learning, explainable artificial intelligence (XAI), Internet of Things (IoT), and blockchain-based approaches for the classification and monitoring of milk quality.

In the literature, studies employing the milk quality grading dataset published on the Kaggle platform have attracted particular attention. chaudhari2025machine compared various machine learning algorithms, including Decision Trees, Random Forests, and Support Vector Machines, to classify milk quality and evaluated model performance using standard classification metrics such as accuracy, precision, recall, and F1 score. Similarly, Bhavsar *et al.* (2023) conducted a detailed analysis of Random Forest and SVM models on the same dataset and demonstrated that Random Forest achieved superior accuracy and generalization performance following appropriate preprocessing and label encoding. Another study Samad *et al.* (2024) compared KNN with its enhanced variant, DW-KNN, for milk quality detection and reported that the proposed method provided improved classification capability. Likewise, Shahzad *et al.* (2025), as well as Çelik (2022), performed accuracy-oriented evaluations by comparing multiple classification algorithms on the Kaggle milk dataset. A common characteristic of these studies is the use of a relatively small but well-defined dataset to demonstrate that milk quality can be automatically classified into low, medium, and high categories.

More recent studies have shifted their focus beyond classification performance to the interpretability of model decision-making processes. Çetintav and Yalçın (2025) performed milk quality classification using Random Forest and HistGradientBoost models on the Kaggle dataset and subsequently integrated XAI techniques such as LIME and Permutation Feature Importance to analyze the contributions of key features, including pH, temperature, and taste, both globally and at the instance level. This approach not only delivered high predictive performance but also provided an explainable framework that transparently revealed the factors underlying milk quality decisions.

At a more advanced methodological level, Kurtanjek (2024) addressed the limitations of correlation-based models by employing causal artificial intelligence approaches. Using Bayesian networks and structural causal modeling, the study investigated causal relationships among factors influencing milk quality, thereby contributing not only to predictive accuracy but also to a deeper understanding of the underlying mechanisms driving quality outcomes. Veena and Poovammal (2025) proposed a hybrid approach combining Principal Component Analysis (PCA) and decision trees with robust scaling to mitigate the effects of outliers. In another study (Tolba *et al.* 2024), a deep learning model integrating GRU and ResNet architectures was evaluated on a milk quality dataset, demonstrating the potential of advanced neural network-based approaches.

Beyond chemical and physicochemical parameters, several studies have addressed milk quality from supply chain and traceability perspectives. Manisha and Jagadeeshwar (2023) introduced a blockchain-supported traceability framework that integrates IoT, blockchain, and deep learning components to ensure transparent and immutable tracking of quality and safety information throughout the food supply chain. Similarly, Goyal *et al.* (2024) developed an AI- and IoT-based multi-sensor system for detecting and classifying milk adulteration, a critical challenge in milk quality assurance; SHAP was employed to enhance the interpretability of model decisions.

Spectroscopic and image-based techniques have also played a significant role in milk quality assessment. One study Thanasirikul *et al.* (2023) utilized color sensor data derived from the rezazurin test to rapidly classify raw milk quality using SVM. Tahtali (2020) examined milk quality based on somatic cell count and compositional parameters in raw milk samples, analyzing the impact of increasing somatic cell levels from a machine learning perspective. Neto *et al.* (2019) achieved high accuracy in detecting milk adulteration and contamination by combining Fourier-transform infrared (FTIR) spectroscopy data with a customized convolutional neural network architecture. Mhapsekar *et al.* (2025) provided a comprehensive review of studies published between 2012 and 2023, comparing machine learning and deep learning approaches for milk quality assessment in terms of data types, methodologies, performance metrics, and emerging research trends.

Despite the extensive literature, studies focusing specifically on the Kaggle milk quality grading dataset that systematically define preprocessing procedures, evaluate multiple machine learning algorithms under fair and consistent conditions, and comprehensively analyze feature importance and decision-making processes within an XAI framework remain limited. This gap highlights the need for more robust and transparent methodological approaches

that achieve both high classification accuracy and strong interpretability in milk quality assessment.

MATERIAL AND METHOD

The methodology adopted in this study is based on the systematic execution of the fundamental stages of the experimental workflow. The process begins with data acquisition, followed by data preprocessing, exploratory analysis of data distributions, and partitioning of the dataset into training and test subsets. After model training and hyperparameter optimization using the training data, the models are evaluated and compared on the test data to identify those with the highest predictive performance. In the final stage, the decision-making mechanisms of the selected models are analyzed using Explainable Artificial Intelligence (XAI) techniques based on LIME and SHAP, and the key features influencing model predictions are examined in detail.

Data Set

In this study, the Milk Quality Dataset published on the Kaggle open data platform was used to develop machine learning models for milk quality classification (Shrijayan (cpluzshrijayan) n.d.). The dataset comprises a total of 1,059 samples, each representing a milk sample described by physical and sensory quality measurements. The data were collected through manual observations by researchers and shared under an open data license ("EU ODP Legal Notice") (Çetintav and Yalçın 2025; Manisha and Jagadeeshwar 2023). The dataset includes seven independent variables that characterize the physical and sensory properties of milk, along with a target variable (Grade) representing overall milk quality:

- pH: Acidity level of milk (numerical; range: 3.0–9.5).
- Temperature: Sample temperature in °C (numerical; range: 34–90).
- Taste: Indicator of taste quality (0 = poor, 1 = good).
- Odor: Indicator of odor quality (0 = poor, 1 = good).
- Fat: Presence of fat content (0 = absent, 1 = present).
- Turbidity: Turbidity level (0 = low, 1 = high).
- Color: Color intensity of the sample (numerical; range: 240–255).

The target variable Grade is categorized into three classes based on sensory and physical criteria: Low (Bad), Medium (Moderate), and High (Good). Although the dataset primarily consists of binary variables, the inclusion of continuous features such as pH, temperature, and color enables the application of both classical machine learning algorithms and explainable artificial intelligence (XAI) techniques [3]. Owing to its open-access nature and the inclusion of parameters widely regarded as critical for milk quality assessment, this dataset provides a suitable and well-established basis for machine learning-based classification and model interpretability studies.

Data Preparation

In this study, a systematic preprocessing pipeline was applied to the dataset to enhance the reliability and generalizability of the machine learning models. In the initial stage, the dataset was inspected for missing values and data type inconsistencies. No missing observations were identified, and all variables were confirmed to have data types suitable for analysis. A detailed examination revealed that a substantial proportion of the samples in the dataset consisted of duplicate records. When these duplicate records were removed, the number of unique samples decreased markedly, which significantly constrained the statistical representativeness of the dataset and the learning capacity of the models. Therefore, to avoid data loss, duplicate records were retained; however, to prevent data leakage, group-based splitting strategies were employed during the model evaluation phase. This strategy ensured that duplicate records corresponding to the same original sample did not appear simultaneously in both the training and test sets. During the modeling process, label encoding was applied to the target variable, Grade. As most of the independent variables are binary, feature scaling was applied exclusively to the continuous variables—pH, Temperature, and Color. These variables were normalized to the [0,1] range using Min–Max normalization. The Min–Max scaling procedure is defined as follows:

$$X_{\text{scaled}} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{1}$$

Here, Xscaled denotes the scaled (normalized) value, X represents the original value, min(X) and max(X) correspond to the minimum and maximum values of the feature across the entire dataset, respectively. This normalization strategy is adopted to prevent differences in feature scales from adversely affecting model performance, particularly for distance-based algorithms such as k-nearest neighbors, support vector machines, and logistic regression.

An examination of class proportions in the dataset revealed no significant class imbalance. The class distributions in the training and test sets are illustrated in Fig. 1. Accordingly, neither oversampling nor undersampling techniques were applied, in order to avoid introducing bias into the model learning process through artificial data manipulation. As a result of these preprocessing steps, the dataset was rendered suitable for training machine learning models and for conducting explainability analyses.

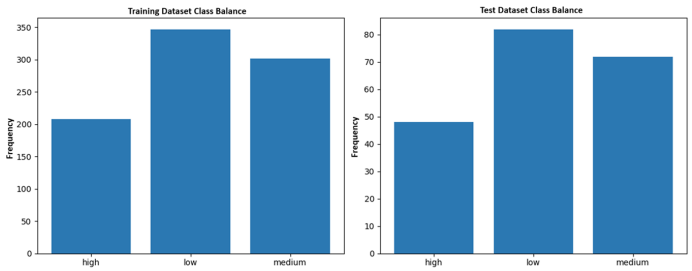

Figure 1 Distribution of class labels in the training and test datasets

Models Preparation

In this study, machine learning models with diverse structural characteristics were evaluated for milk quality classification. Within the scope of the comparative analysis, RF, KNN, SVC, XGB, LGBM, and ANN models were employed. Following an initial assessment using baseline hyperparameter settings, hyperparameter optimization was conducted to enhance model performance. GridSearchCV was adopted as a conventional baseline approach for optimization within limited and discrete hyperparameter spaces. However, due to the high computational cost and the exhaustive nature of grid-based search particularly in continuous and high-dimensional parameter spaces a more flexible and computationally efficient optimization strategy was required.

Accordingly, Particle Swarm Optimization (PSO)-based hyperparameter tuning was applied, particularly to tree-based models

(RF, XGB, and LGBM) and ANN architectures. PSO is a population-based heuristic optimization algorithm that facilitates rapid convergence toward global optima in complex and continuous search spaces. By iteratively updating candidate solutions (particles) based on both individual best positions and the global best solution, PSO enables a more efficient exploration of the hyperparameter space compared to grid-based methods. During the hyperparameter optimization process, both accuracy and macro-averaged F1 score were jointly considered as performance metrics. While accuracy reflects overall classification performance, the macro-F1 score assigns equal importance to each class, thereby mitigating potential biases arising from class distribution differences. All optimization procedures were conducted in accordance with the group-based data splitting strategy implemented to prevent data leakage caused by repeated samples. The optimal hyperparameter configurations obtained for each model are reported in Table 1.

Evaluation Criteria

To comprehensively and reliably evaluate the performance of the classification models developed in this study, four widely used performance metrics were considered: accuracy, precision, recall, and F1 score. Accuracy represents the overall performance of a model and is defined as the ratio of correctly classified instances to the total number of instances. However, accuracy alone may be insufficient in scenarios involving class imbalance or when certain classes are of greater importance. Precision evaluates the extent of false positive predictions by indicating the proportion of instances predicted as positive that are actually correct. Recall, in contrast, captures the impact of false negatives by measuring the proportion of true positive instances that are correctly identified. The F1 score, which provides a balance between precision and recall, is defined as the harmonic mean of these two metrics and enables a more robust comparison of model performance across classes (Chaudhari *et al.* 2025; Horasan *et al.* 2019). The mathematical definitions of these performance metrics are provided in Equations (2)–(5).

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Instances} \tag{2}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{3}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{4}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

In addition to these quantitative metrics, a confusion matrix was employed to analyze model performance on a per-class basis in greater detail. The confusion matrix offers a visual representation of correct and incorrect predictions for each class, thereby clearly illustrating the strengths and weaknesses of the models with respect to individual class predictions.

Explainable Artificail Intelligent

In this study, the classification model that achieved the highest performance during the evaluation was selected for the analysis of the explainability. To enhance the transparency and interpretability of the decision-making mechanism of the selected model, two Explainable Artificial Intelligence (XAI) techniques, LIME and SHAP were jointly employed.

Due to its model-agnostic nature, the LIME method constructs a local linear approximation around a specific instance, thereby revealing which features influence individual predictions and in which direction. This capability provides a substantial advantage, particularly in interpreting and justifying instance-level predictions and in understanding the behavior of the model in specific samples (Ribeiro *et al.* 2016; Chowdhury *et al.* 2024).

The SHAP method, in contrast, leverages Shapley values derived from cooperative game theory to quantitatively estimate the contribution of each feature to the model output, offering explanations at the global level. This enables a systematic assessment of the variables that shape the overall decision structure of the model and their relative importance throughout the dataset (Sermmany *et al.* 2024; Shapley 1953).

By combining LIME and SHAP, it becomes possible to explain both local (instance-level) and global (model-level) behaviors of the selected classifier. Consequently, instead of simply reporting high performance metrics, the decision-making processes of the model are examined within a comprehensive and scientifically grounded explainability framework, thus substantially strengthening the interpretability of the proposed approach.

EXPERIMENTAL RESULTS

Models Performance

In this study, various machine learning models were evaluated for milk quality classification, and their performance was compared using accuracy, macro-precision, macro-recall, and macro-F1 score metrics. The obtained results are summarized in Table 2.

An examination of Table II shows that the XGB and LGBM models optimized using PSO achieved the highest performance across all evaluation metrics. In particular, the XGB(PSO) model emerged as a representative high-performing model, achieving a test accuracy of 89.1 % and a Macro-F1 score of 0.8777. These findings indicate that boosting-based methods are capable of effectively modeling the complex and non-linear relationships inherent in milk quality data.

Although RF(PSO) and ANN(PSO) demonstrated moderate performance, the LR, KNN, and SVC models optimized via Grid Search showed notably lower results in comparison. The relatively low Macro-F1 scores of these models suggest limitations in maintaining balanced performance across classes. Overall, the results highlight that PSO-based hyperparameter optimization substantially enhances classification performance, particularly for ensemble and boosting-based models.

Fig. 2 illustrates the confusion matrices obtained on the test dataset for the PSO-optimized XGB, RF, and ANN models. The XGB(PSO) model correctly classifies the high and low classes without error, whereas a limited number of medium class instances are misclassified as high. In contrast, the RF(PSO) and ANN(PSO) models exhibit a more pronounced confusion between the medium and high classes. These observations further support the superior inter-class discrimination capability of the XGB model and are consistent with the higher Macro-F1 score reported in Table 2.

XAI Results

To enhance the transparency of the proposed classification framework, explainable artificial intelligence (XAI) techniques were employed to analyze both global and local decision mechanisms of the optimized XGBoost model. In this context, SHAP was adopted as the primary explanation method, while LIME was used to provide complementary local verification.

Table 1 est Hyperparameters and Optımızatıon Methods Selected for the Compared Models

Model	Optimization	Best hyperparameters
Logistic Regression	Grid Search	C=100, solver=lbfgs
KNN	Grid Search	n_neighbors=3, p=2, weights=distance
SVM (RBF)	Grid Search	C=100, gamma=scale, kernel=rbf
Random Forest	PSO	n_estimators=353, max_depth=14, min_samples_split=2, min_samples_leaf=1
XGBoost	PSO	n_estimators=588, max_depth=9, learning_rate≈0.243, subsample≈0.675, colsample_bytree≈0.957, reg_lambda≈2.70
LightGBM	PSO	n_estimators=231, num_leaves=104, learning_rate≈0.295, min_child_samples=31, subsample≈0.657, colsample_bytree≈0.913
ANN (MLP)	PSO	hidden_layer_sizes=(87,), alpha≈ 6.1×10^{-4}, learning_rate_init≈0.0118

Table 2 Performance Table

Model	Test_Accuracy	Test_MacroPrecision	Test_MacroRecall	Test_MacroF1
XGB(PSO)	0.891089	0.895238	0.898148	0.877744
LGBM(PSO)	0.891089	0.895238	0.898148	0.877744
RF(PSO)	0.811881	0.852713	0.845528	0.804944
ANN(PSO)	0.782178	0.834028	0.807588	0.781656
LR(Grid)	0.594059	0.766026	0.653117	0.596633
KNN(Grid)	0.594059	0.766026	0.653117	0.596633
SVC(Grid)	0.594059	0.691135	0.653117	0.581848

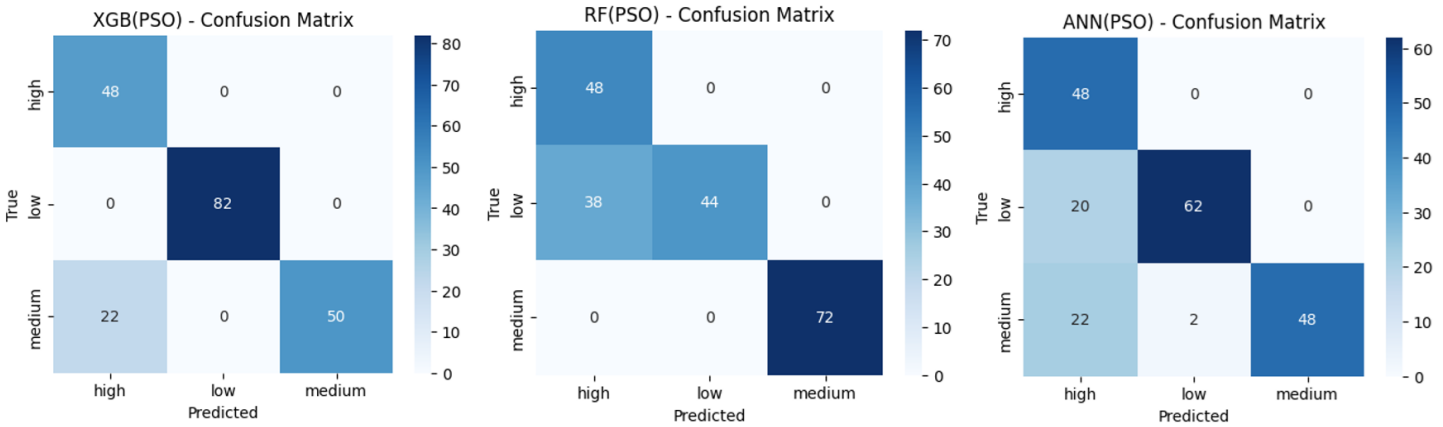

Figure 2 Confusion matrices obtained on the test dataset for the PSO-optimized XGB, RF, and ANN models.

Figure 3 presents the global feature importance derived from mean absolute SHAP values. The results indicate that pH, Fat, and Temperature are the most influential variables in the model's decision process, followed by Odor, Colour, Turbidity, and Taste. This ranking demonstrates that physicochemical properties of milk, particularly acidity level and fat content, play a dominant role in quality grade discrimination. Compared to discrete sensory-related attributes, continuous physicochemical features exhibit

stronger and more stable contributions.

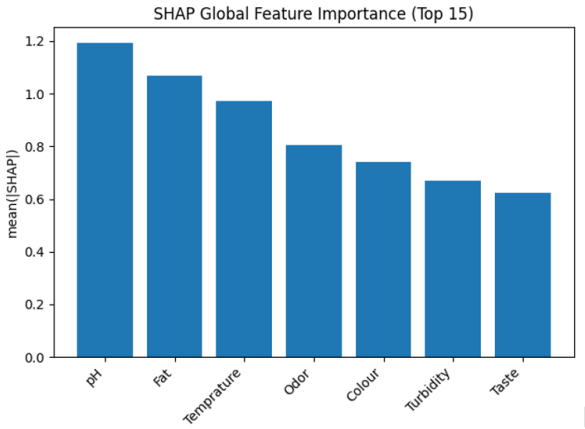

Figure 3 SHAP Global Feature Importance Based on Mean Absolute SHAP Values.

While global importance rankings reveal which features matter most, they do not capture how feature values influence class predictions. To address this, Figure 4 illustrates the SHAP summary (beeswarm) plot for the most frequently observed class. In this visualization, each point represents an individual sample, colored according to its feature value. Positive SHAP values indicate an increased likelihood of the target class, whereas negative values reduce the corresponding class score. The plot reveals that pH, Fat, and Temperature not only have high importance but also exhibit bidirectional effects, depending on their observed values. This behavior highlights their role in defining class boundaries rather than acting as monotonic predictors. These findings underline that milk quality assessment is governed by complex feature interactions, where identical variables may support different class outcomes under varying conditions.

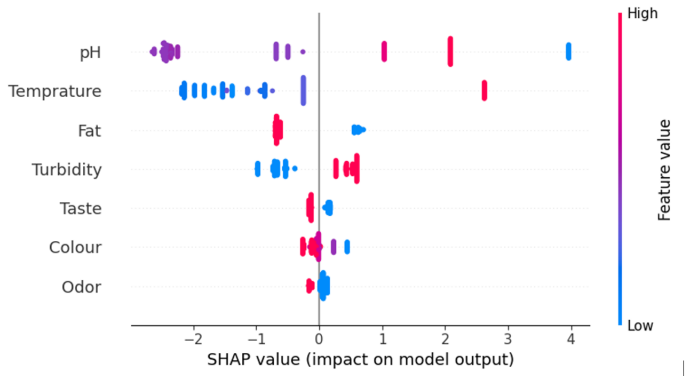

Figure 4 SHAP Summary (Beeswarm) Plot for the Most Frequent Class.

To further investigate the model's decision-making process at the instance level, a local SHAP analysis was conducted on a misclassified test sample. Figure 5 shows the SHAP waterfall plot for an instance whose true label was medium but was predicted as high. The visualization reveals that Fat, Colour, Taste, and Temperature contributed positively toward the high class prediction, while pH and Odor exerted negative influence. Despite the presence of counteracting signals, the cumulative positive contributions outweighed the negative effects, resulting in an incorrect classification.

This example demonstrates how borderline samples located near class boundaries may be sensitive to competing feature contributions, leading to misclassification even in high-performing models.

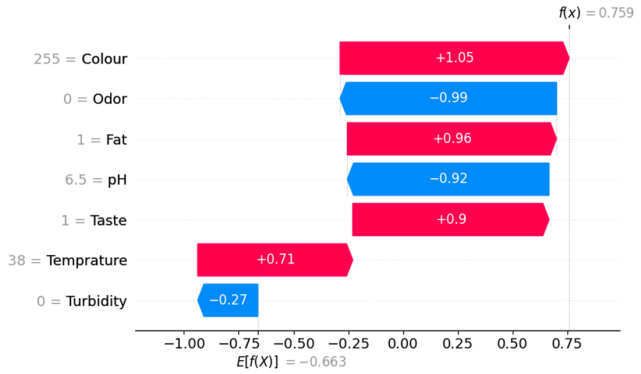

Figure 5 Local SHAP Explanation of a Misclassified Test Instance.

To validate the reliability of the SHAP-based interpretations, a LIME local explanation was generated for the same test instance. Figure 6 presents the LIME explanation for the predicted high class. Consistent with the SHAP analysis, LIME identified Fat, Colour, Taste, and Temperature as positive contributors to the high class, while pH and Odor acted as negative factors. The agreement between SHAP and LIME reinforces the robustness of the observed explanations and confirms that the model's local decision logic is not dependent on a single interpretability method.

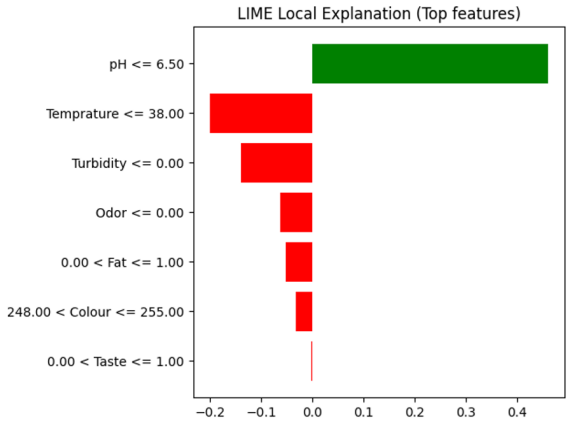

Figure 6 LIME Local Explanation for the Predicted High Class.

Overall, the combined global and local XAI analyses demonstrate that the proposed model relies on meaningful and domain-consistent features when assessing milk quality. The use of group-based data splitting ensures that these explanations are not affected by data leakage, thereby increasing their reliability. Moreover, the consistency between SHAP and LIME interpretations supports the transparency and trustworthiness of the model, particularly in borderline cases where classification uncertainty is inherently higher.

COMPARISON WITH RELATED STUDIES

When compared with existing studies in the literature on milk quality classification, the approach proposed in this study exhibits notable differences, particularly with respect to data preprocessing

strategies, and highlights several common methodological gaps. The vast majority of the reviewed studies rely on the same publicly available dataset obtained from Kaggle (1,059 samples with 7 features) and adopt similar preprocessing assumptions, largely due to the absence of missing values in the dataset. However, the presence of duplicate records an issue that is critical for data integrity and the reliability of model evaluation and the manner in which such records are handled are not explicitly reported in most studies (Bhavsar *et al.* 2023; Çetintav and Yalçın 2025; Shahzad *et al.* 2025; Çelik 2022; Mu *et al.* 2020; Samad *et al.* 2024; Kurtanjek 2024). Only Study 1 reports the identification and direct removal of duplicate samples from the dataset.

The analyses conducted in the present study revealed a substantial number of duplicate records within the dataset. It was determined that directly eliminating these records could significantly reduce the representativeness of the data and adversely affect the learning capacity of the models. To address this issue, group-based data partitioning strategies were employed to prevent data leakage while preserving the full set of available samples. This strategy ensured that duplicate instances originating from the same source did not appear simultaneously in both the training and test sets, thereby enabling a more realistic and reliable evaluation of model performance. To the best of our knowledge, none of the existing studies in the literature explicitly address duplicate data management at this level.

With respect to feature scaling, Min–Max normalization is commonly adopted in the literature (Bhavsar *et al.* 2023; Çetintav and Yalçın 2025; Shahzad *et al.* 2025; Samad *et al.* 2024), while StandardScaler is preferred in (Chaudhari *et al.* 2025; Tolba *et al.* 2024), and (Kumari *et al.* 2023), Robust Scaling is applied only in (Veena and Poovammal 2025). In contrast, the present study adopts a targeted preprocessing strategy in which scaling is not applied to binary variables, and Min–Max normalization is used exclusively for continuous features such as pH, temperature, and color. This selective approach aims to improve the performance of distance-based algorithms while avoiding unnecessary transformations of discrete sensory-related attributes.

Different strategies have also been reported in the literature to address potential class imbalance. While SMOTE is employed in (Tolba *et al.* 2024), balance is achieved through strategic sample selection in (Çelik 2022), and (Çetintav and Yalçın 2025) reports that the dataset is inherently balanced. In the distribution analyses conducted in this study, no significant class imbalance was detected. Consequently, oversampling or undersampling techniques were not applied in order to avoid introducing additional bias through artificial manipulation of the data distribution.

A systematic comparison of the proposed methodology with related studies is provided in Table 3. As summarized in the table, most existing studies do not explicitly report how duplicate records are handled, nor do they apply selective scaling strategies or group-based data partitioning. In contrast, the proposed approach integrates these steps in a unified preprocessing framework, highlighting its methodological distinctiveness.

When the accuracy comparisons presented in Table 4 are examined, it becomes evident that many of the high performance values reported in the literature are obtained without explicitly addressing duplicate data management or potential data leakage risks. In contrast, the cautious and controlled preprocessing strategy adopted in this study results in lower accuracy values for some models; however, these results are considered to more accurately reflect real-world generalization performance, as they are obtained under strict data leakage control conditions.

Overall, this study introduces a minimalist and reliability-oriented preprocessing philosophy by prioritizing duplicate-aware data partitioning, avoiding unnecessary scaling operations, and refraining from artificial data balancing techniques. Future studies may further investigate the origins of duplicate samples (e.g., repeated measurements or data acquisition artifacts) and systematically evaluate the impact of different normalization strategies on model stability and robustness.

CONCLUSION

This study comparatively evaluates the classification performance of various supervised machine learning algorithms using an open-access milk quality dataset. The primary objective of the study is to move beyond approaches that focus solely on high accuracy values and to present a holistic methodological framework that jointly considers the reliability of data preprocessing, the validity of model evaluation, and the interpretability of decision-making processes. The findings indicate that, particularly in datasets containing repeated samples derived from the same source, data leakage can artificially inflate model performance. Accordingly, instead of directly removing duplicate records, group-based data partitioning strategies were employed to prevent data loss while preserving the models' ability to generalize to unseen data. Through this approach, models were evaluated under more realistic conditions, and attention was drawn to an important methodological issue that is frequently overlooked in the literature.

During the data preprocessing stage, scaling was applied exclusively to the necessary continuous variables, and artificial resampling techniques were deliberately avoided for a dataset that exhibited no class imbalance. This targeted and controlled data preparation strategy highlights that, especially for datasets with a high proportion of duplicate samples, detailed data analysis and informed preprocessing decisions are as critical as the choice of learning algorithm itself. The modeling results demonstrate that tree-based and boosting-based models optimized using Particle Swarm Optimization (PSO) achieve more balanced and consistent performance in milk quality classification. These findings suggest that PSO-based hyperparameter optimization provides a more efficient and effective search process in high-dimensional and continuous parameter spaces compared to traditional grid-based approaches.

To move beyond predictive performance, explainable artificial intelligence (XAI) techniques were employed to analyze the decision-making processes of the models. Using LIME and SHAP, physicochemical attributes such as pH, fat content, and temperature were identified as key determinants in milk quality prediction. This analysis enables model outputs to be evaluated not only through numerical performance metrics but also in an interpretable and well-justified manner. In conclusion, this study presents a reliable, transparent, and methodologically sound machine learning framework for milk quality classification by jointly addressing duplicate data management, data leakage prevention, targeted data preprocessing, hyperparameter optimization, and explainability analysis.

Future studies aim to further investigate the impact of explainable AI analyses across different model architectures and data partitioning strategies, as well as to evaluate the contribution of duplicate-aware data management approaches to generalization performance on larger and more diverse milk quality datasets. In addition, future work may explore the integration of domain knowledge and sensor-level metadata to further enhance interpretability and robustness.

Table 3 Comparison of the Proposed Approach with Related Studies in the Literature

Articles	Missing Value Check	Scaling Method	Duplicate Handling	Dimension Reduction	Class Imbalance Analysis
(Chaudhari *et al.* 2025)	+	StandardScaler	+(removed, impact of data reduction not reported)	-	-
(Bhavsar *et al.* 2023)	+	Min–Max	-	-	-
(Çetintav and Yalçın 2025)	+	Min–Max	-	-	+(balanced)
(Samad *et al.* 2024)	+	Min–Max	-	-	+(balanced)
(Veena and Poovammal 2025)	+	Robust Scaling	-	+ (PCA)	-
(Tolba *et al.* 2024)	+	StandardScaler	-	-	-(imbalanced – SMOTE)
(Kumari *et al.* 2023)	+	StandardScaler	+(no duplicate values present)	+ (PCA)	-
Ours	+	Min–Max(continuous only)	+(group-based, leakage-aware)	-	+(balanced, no resampling)

Table 4 Accuracy Comparison with Related Studies in the Literature

Model \ Articles	(Chaudhari *et al.* 2025)	Bhavsar *et al.* (2023)	(Çetintav and Yalçın 2025)	(Samad *et al.* 2024)	(Veena and Poovammal 2025)	(Tolba *et al.* 2024)	(Kumari *et al.* 2023)	**Ours**
Logistic Regression	0.36	-	-	-	0.60	-	0.8490	0.5941
Random Forest	0.99	0.92	0.995	-	0.75	0.914	0.9968	0.8119
SVM	-	0.57	0.566	-	0.75	0.868	0.9528	0.5941
KNN	0.98	-	0.985	0.985	-	-	0.9968	0.5941
ANN / MLP	0.45	-	-	-	-	0.933	-	0.7822
XGBoost	-	-	0.973	-	-	-	-	0.8911
LightGBM	-	-	0.970	-	-	-	-	0.8911

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

Arrighi, L., I. A. de Moraes, M. Zullich, M. Simonato, D. F. Barbin, *et al.*, 2025 Explainable artificial intelligence techniques for interpretation of food datasets: a review. arXiv preprint arXiv:2504.10527 .

Azad, T. and S. Ahmed, 2016 Common milk adulteration and their detection techniques. International Journal of Food Contamination **3**: 22.

Bhavsar, D., Y. Jobanputra, N. K. Swain, and D. Swain, 2023 Milk quality prediction using machine learning. EAI Endorsed Transactions on Internet of Things **10**.

Çelik, A., 2022 Using machine learning algorithms to detect milk quality. Eurasian Journal of Food Science and Technology **6**: 76–87.

Çetintav, B. and A. Yalçın, 2025 Explainable machine learning framework for milk quality grading. Kocatepe Veterinary Journal **18**: 227–235.

Chaudhari, A., R. Mane, A. Khot, A. Kadam, and N. Rajam, 2025 Machine learning-based classification for milk quality assessment. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pp. 1441–1446, IEEE.

Chowdhury, R., R. Das, F. B. F. Ananna, A. Saha, S. Nawar, *et al.*, 2024 Unveiling predictive factors in apple quality: Leveraging

lime, shap, and the synergy of machine learning models and artificial neural networks. In *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, pp. 1026–1031, IEEE.

Goyal, K., P. Kumar, and K. Verma, 2024 Xai-empowered iot multisensor system for real-time milk adulteration detection. Food Control **164**: 110495.

Horasan, F., H. Erbay, F. Varçın, and E. Deniz, 2019 Alternate low-rank matrix approximation in latent semantic analysis. Scientific Programming **2019**: 1095643.

Kumari, S., M. K. Gourisaria, H. Das, and D. Banik, 2023 Deep learning based approach for milk quality prediction. In *2023 11th International conference on emerging trends in engineering & technology-signal and information processing (ICETET-SIP)*, pp. 1–6, IEEE.

Kurtanjek, Ž., 2024 Causal artificial intelligence models of food quality data. Food Technology and Biotechnology **62**: 102–109.

Manisha, N. and M. Jagadeeshwar, 2023 Bc driven iot-based food quality traceability system for dairy product using deep learning model. High-Confidence Computing **3**: 100121.

Mhapsekar, R., D. Kilbane, S. Davy, L. Abraham, M. Fenelon, *et al.*, 2025 A systematic review of the internet of things and artificial intelligence applications in milk quality monitoring and analysis. International Journal of Dairy Technology **78**: e70049.

Mu, F., Y. Gu, J. Zhang, and L. Zhang, 2020 Milk source identification and milk quality estimation using an electronic nose and machine learning techniques. Sensors **20**: 4238.

Murphy, S. C., N. H. Martin, D. M. Barbano, and M. Wiedmann, 2016 Influence of raw milk quality on processed dairy products: How do raw milk quality test results relate to product quality and yield? Journal of Dairy Science **99**: 10128–10149.

Neto, H. A., W. L. Tavares, D. C. Ribeiro, R. C. Alves, L. M. Fonseca, *et al.*, 2019 On the utilization of deep and ensemble learning to detect milk adulteration. BioData Mining **12**: 13.

Peri, C., 2006 The universe of food quality. Food quality and preference **17**: 3–8.

Polat, O., S. G. Akçok, M. A. Akbay, D. Topaloğlu, S. Arslan, *et al.*, 2021 Classification of raw cow milk using information fusion framework. Journal of Food Measurement and Characterization **15**: 5113–5130.

Ribeiro, M. T., S. Singh, and C. Guestrin, 2016 " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Samad, A., S. Taze, and M. K. Uçar, 2024 Enhancing milk quality detection with machine learning: A comparative analysis of knn and distance-weighted knn algorithms. Int. J. Innov. Sci. Res. Technol **9**: 2021–2029.

Sarveswaran, S., S. Jha, B. Soundarya, *et al.*, 2023 Milksafe: a hardware-enabled milk quality prediction using machine learning. In *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, pp. 1–6, IEEE.

Sermmany, K., P. Wanjantuk, and W. Leelapatra, 2024 Utilizing explainable artificial intelligence (xai) to identify determinants of coffee quality. In *2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 696–703, IEEE.

Shahzad, A., S. Javaid, and Z. Alamsyah, 2025 Milk quality detection using machine learning. Engineering Proceedings **107**: 119.

Shapley, L., 1953 Stochastic gamesproceedings of the national academy of sciences of the usa 39, 1095–1100 (chapter 1 in this volume). MathSciNet zbMATH .

Shrijayan (cpluzshrijayan), n.d. Milk Quality Prediction Dataset. https://www.kaggle.com/datasets/cpluzshrijayan/milkquality, Accessed: 2026-01-25.

Sunithamani, S., D. Muralidhar, G. Anne, and C. N. Sruthi, 2024 Milk quality prediction using machine learning integrated with arduino. In *2024 10th International Conference on Communication and Signal Processing (ICCSP)*, pp. 1268–1273, IEEE.

Tahtali, Y., 2020 Classification of raw milk composition and somatic cell count in water buffaloes with support vector machines. Kafkas Üniversitesi Veteriner Fakültesi Dergisi **26**.

Thanasirikul, C., A. Patumvan, D. Lipsky, S. Bovonsombut, P. Singjai, *et al.*, 2023 Rapid assessment and prediction of microbiological quality of raw milk using machine learning based on rgb-colourimetric resazurin assay. International Dairy Journal **146**: 105750.

Tolba, A., N. Mostafa, A. Mohamed, and K. Sallam, 2024 Hybrid deep learning approach for milk quality prediction. Precis. Livest. J. **1**: 1–13.

Veena, V. and E. Poovammal, 2025 An improved multi classification of milk quality using machine learning. In *2025 2nd International Conference on Trends in Engineering Systems and Technologies (ICTEST)*, volume 1, pp. 1–6, IEEE.

How to cite this article: Çetinkaya, Z., Horasan, F., Aydilek, H., and Erten, M. Y. Predictive Modeling for Milk Quality Using Machine Learning and XAI Algorithms. *ADBA Computer Science*, 3(1), 63-71, 2026.