Vol. 9 (2025): Proceedings of the 2025 International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability
(IC-AIMEES 2025)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.202501

Machine Learning-based Milk Quality Prediction

Bojun Zhao*

Applied Mathematics, Beijing Normal-Hongkong Baptist University, Zhuhai, China

**Corresponding author: Bojun Zhao*

Abstract

With improvements in people's living standards, food safety has gradually received increasing attention. Milk quality is an important aspect of food safety. By classifying quality, we can obtain milk of different grades for consumers to choose from. This report applies the MLP, logistic regression, SVM and XGBoost algorithms to determine a suitable model through comparison and decision-making.

Keywords

milk, food quality, multilayer perceptron, XGBoost, SVM, logistic regression

1. Background Motivation

Milk is a globally essential nutritional commodity that serves as a primary source of calcium, protein, and vitamins for billions of people. Its quality directly impacts consumer health, product shelf life, and the reputation of dairy industries. Traditional methods for assessing milk quality rely on laboratory tests, sensory evaluations, and chemical analyses, which are time-consuming, labor-intensive, and often impractical for real-time decision-making in large-scale production environments. Many indicators affect the quality of milk, such as pH, temperature, turbidity, taste, odor, and color. Each indicator reflects whether there is an issue with the milk in a specific aspect. We cannot clearly understand multiple data points simultaneously through conventional methods, so we need to establish an evaluation mechanism to classify milk quality. Machine learning is an efficient approach—by training on a dataset where the quality has already been labeled, we can classify the grades of the remaining milk, thereby providing consumers with a clearer understanding of the product's condition. In this project, we employ logistic regression, SVM, MLP, and XGBoost algorithms for training. By comparing the models' performance, we aim to identify the optimal solution for this problem.

2. Data Processing

2.1 Related Works

Early work relied on statistical methods such as logistic regression and discriminant analysis to correlate milk quality with measurable features. For example, a study by demonstrated that pH and temperature are strong predictors of spoilage, achieving 78% accuracy via logistic regression. Modern approaches leverage supervised learning algorithms. Research by applied **Random Forests** to a dataset of 500 milk samples, achieving 89% accuracy in classifying milk into *high*, *medium*, and *low* grades. Similarly, support vector machines (SVMs) were employed by to handle imbalanced data, although computational complexity remains

a challenge. Convolutional neural networks (CNNs) have been tested for visual quality inspection (e.g., color analysis). For example, a study using milk color values (RGBs) achieved 92% accuracy in detecting spoilage. However, these models demand large labeled datasets and lack interpretability.

2.2 Dataset

The dataset was constructed for dairy product quality classification research, comprising a total of 1,059 complete data records with 8 key feature dimensions. The specific fields include pH value, temperature, taste, odor, fat content, turbidity, color value, and the target variable quality grade.

In terms of data types, the pH value, temperature, and color value are continuous numerical features that represent the acidity, storage environment temperature, and visual appearance of dairy products, respectively. The numerical types for taste, odor, fat content, and turbidity require further verification on the basis of actual measurement methods. The quality grade, as the classification target, is explicitly divided into three ordered categories: “high,” “low,” and “medium.”

For data partitioning, the dataset employs a stratified sampling strategy to ensure balanced class distributions in both the training set (1,024 records) and the test set (211 records). As shown in Figure 1, in the training set, the “low” grade samples account for the highest proportion (~34.2%), whereas the test set exhibits a more balanced distribution: “high” (51 records, 24.2%), “low” (86 records, 40.8%), and “medium” (75 records, 35.5%). This scientifically designed partitioning not only ensures sufficient training data but also guarantees that the test set comprehensively evaluates the model's classification performance across different quality grades, providing a reliable data foundation for subsequent modeling work.

Figure 1: Distribution of the number of different categories in the training and validation sets: (a) training set and (b) validation set

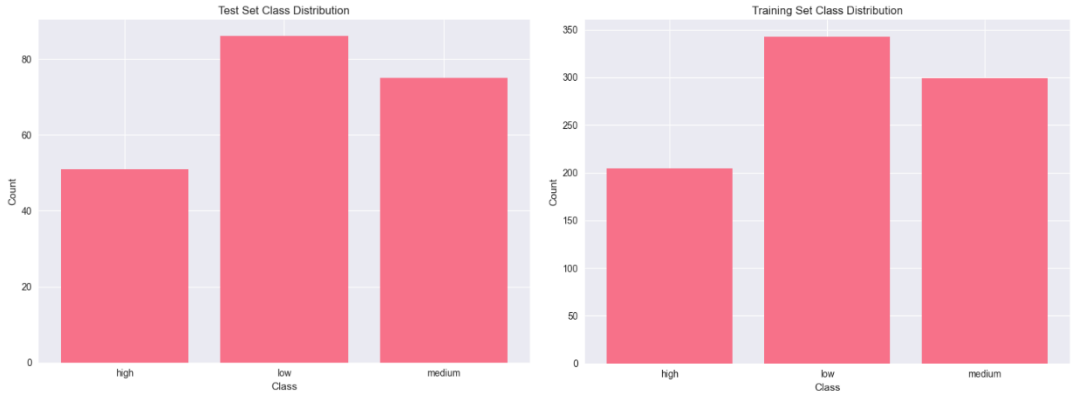
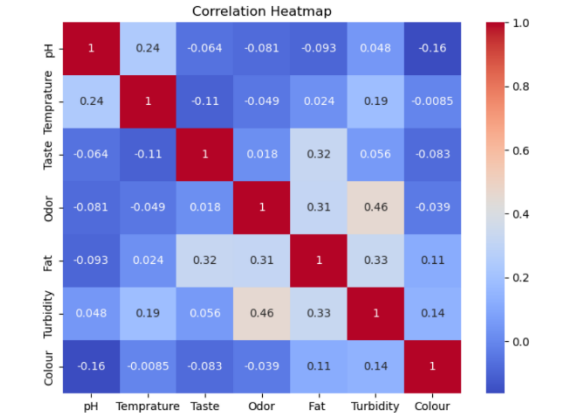


Figure 2: Correlation Heatmap and Descriptive Statistics



	pH	Temprature	Taste	Odor	Fat	Turbidity	Colour
count	1059.000000	1059.000000	1059.000000	1059.000000	1059.000000	1059.000000	1059.000000
mean	6.630123	44.226629	0.546742	0.432483	0.671388	0.491029	251.840415
std	1.399679	10.098364	0.498046	0.495655	0.469930	0.500156	4.307424
min	3.000000	34.000000	0.000000	0.000000	0.000000	0.000000	240.000000
25%	6.500000	38.000000	0.000000	0.000000	0.000000	0.000000	250.000000
50%	6.700000	41.000000	1.000000	0.000000	1.000000	0.000000	255.000000
75%	6.800000	45.000000	1.000000	1.000000	1.000000	1.000000	255.000000
max	9.500000	90.000000	1.000000	1.000000	1.000000	1.000000	255.000000

3. Methodology

3.1 Logistic Regression

Logistic regression (LR) is a statistical classification model that estimates the probability of a discrete outcome on the basis of input features. Despite its simplicity, it is a powerful baseline for classification problems and offers strong interpretability, especially when feature relationships are approximately linear.

In multiclass scenarios, logistic regression is typically extended via the one-vs-rest (OvR) or multinomial approach. It calculates class membership probabilities via the Softmax function, which maps input features to a normalized probability distribution over all classes. Each weight coefficient in LR directly reflects the influence of a specific feature on a given class, making it ideal for understanding feature importance (Das, 2023).

Given its linear nature, LR may underperform in capturing complex nonlinear relationships present in the milk dataset. However, it is valuable for benchmarking model performance and for identifying feature relevance with high transparency.

3.2 Support Vector Machine (SVM)

The support vector machine (SVM) is a supervised learning algorithm that is particularly well suited for classification tasks, especially when dealing with high-dimensional and small-to-medium-sized datasets. The primary objective of SVM is to find an optimal hyperplane that separates classes in the feature space with the maximum margin, thereby ensuring strong generalization performance.

SVM operates under two main principles: (1) maximizing the margin between support vectors—the data points that lie closest to the decision boundary—and (2) minimizing classification error by introducing a soft margin with a penalty term for misclassified points. For datasets that are not linearly separable, the SVM employs the kernel trick to project data into a higher-dimensional space where linear separation is feasible. Common kernel functions include linear, polynomial, and radial basis function (RBF) functions, with the RBF kernel being particularly effective for nonlinear problems such as milk quality classification (Ding et al., 2011).

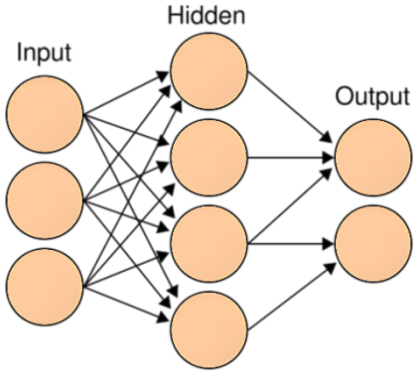
In this study, SVM is selected for its robust generalization ability and effectiveness in handling feature interactions in the milk dataset. By leveraging the RBF kernel, the SVM can model complex nonlinear relationships among features such as pH, turbidity, and fat content.

3.3 Multilayer Perceptron

The multilayer perceptron (MLP) is a classical feedforward artificial neural network that serves as one of the foundational models in deep learning and is widely applied in supervised learning tasks such as classification and regression. It is particularly well suited for processing structured data and performing feature extraction and pattern recognition in image data. A defining characteristic of the MLP is its unidirectional information flow from input to output, which is devoid of feedback connections and ensures structural simplicity and computational stability. Owing to its robust nonlinear modeling capability, the MLP demonstrates exceptional performance in pattern recognition and data modeling.

As shown below, the architecture of an MLP consists of three fundamental components: (1) Input layer: This layer receives raw input data, with the number of neurons corresponding to the dimensionality of the input features, ensuring complete data propagation into the network. (2) Hidden layers: Comprising one or multiple layers of neurons, these layers extract complex patterns from the input data through nonlinear activation functions (e.g., ReLU, sigmoid, and tanh). Each neuron in a hidden layer is fully connected to all the neurons in the subsequent layer, enabling progressive feature abstraction and transformation. (3) Output Layer: This layer produces the final prediction, generating either class labels (for classification tasks) or continuous values (for regression tasks). The number of neurons in the output layer is directly determined by the task objective.

Figure 3: MLP structure



During forward propagation, the input data undergo sequential processing through weighted summation and nonlinear activation functions across each layer. This hierarchical transformation progressively abstracts low-level input features into higher-level representations, ultimately mapping them to the desired output space. Consider an L-layer MLP (including the input, hidden, and output layers), where the input feature dimension is n and the number of neurons in the l layer is denoted as m_l (where $l \in \{1..., L\}$).

The input layer directly receives the raw input data, typically represented as a vector $\mathbf{x} \in \mathbb{R}^n$, where n denotes the dimensionality of the input features. Unlike other layers, the input layer performs no computational transformations; it solely serves to propagate the data to the first hidden layer:

$$a^{(0)} = \mathbf{x} \tag{1}$$

For each hidden layer (denoted as the l-th layer, where $l = 1, 2..., L-1$), the computation involves two key stages: (1) linear transformation and (2) nonlinear activation. The layer receives input $a^{(l-1)}$ from the preceding layer (or the input layer when l=1) and processes it as follows:

$$z^{(L)} = W^{(L)} a^{(L-1)} + b^{(L)} \tag{2}$$

$$a^{(L)} = \sigma_{\text{out}}\left(z^{(L)}\right) \tag{3}$$

$$\text{Softmax}\left(z^{(L)}\right)_i = \frac{e^{z_i^{(L)}}}{\sum_{j=1}^{m^{(L)}} e^{z_j^{(L)}}} \tag{4}$$

Finally, the output of the last layer $a^{(L)}$ serves as the model's prediction, which is compared with the ground-truth labels to compute the loss function.

3.4 XGBoost

XGBoost is a highly efficient and scalable gradient boosting machine learning algorithm widely applied to supervised learning tasks, including classification, regression, and ranking. Built upon the gradient boosting decision tree (GBDT) framework, it constructs a powerful learner by integrating multiple weak learners (typically decision trees) to achieve high-precision predictions (Chen and Guestrin, 2016).

XGBoost's core principle lies in iteratively adding decision trees to minimize the objective loss function. Each tree is optimized on the basis of prediction residuals from previous trees, progressively refining model predictions. The algorithm introduces regularization terms to control model complexity and accelerates optimization through second-order gradient information (Hessian matrix). Its main advantages are as follows: (1) Objective function optimization:

Both the loss function and regularization terms are integrated to construct a composite objective function. (2) Second-order gradient information: accelerated gradient descent using the second-order derivative of the loss function (Hessian). (3) Parallelization and optimization: These methods support feature parallelism, data chunking and distributed computing to improve training speed.

The training process of XGBoost can be summarized as follows: initialize the model, iteratively add the decision tree, adjust the tree structure and weights by optimizing the objective function, and finally output the integrated prediction results.

4. Experimental

4.1 Algorithm Setup

4.1.1 Logistic Regression Setup

Logistic regression is a generalized linear model widely used in classification problems, and its basic principle is to represent the probability of the occurrence of an event by introducing a sigmoid function that maps the continuous values of the output of a linear regression model to an interval between 0 and 1. In terms of parameter estimation, logistic regression usually employs great likelihood estimation by constructing a log-likelihood function and iteratively optimizing the parameters to maximize the probability of the sample data occurring under the current model. The objective function of the model is nonlinear and usually needs to be solved via numerical optimization methods such as gradient descent or the simulated Newton method. Logistic regression essentially models a weighted linear combination of input features and achieves nonlinear mapping in the probabilistic sense with the help of a sigmoid function; therefore, it is highly interpretable. Its decision boundary is a linear hyperplane, which is suitable for the case where the relationship between features and labels is linearly differentiable or approximately linear.

4.1.2 SVM Setup

SVM solves for the best hyperplane by optimizing a convex quadratic programming problem, which involves minimizing the complexity of the model (i.e., minimizing the sum of squares of the weights) while limiting the misclassification of the training samples. This optimization problem can be solved via the Lagrange multiplier method. For the nonlinearly differentiable case, the SVM can map the input features to a high-dimensional space through a kernel function (kernel function), which makes the originally linearly indistinguishable data linearly differentiable in the high-dimensional space. Commonly used kernel functions include linear kernels, polynomial kernels, Gaussian kernels and so on.

4.1.3 Multilayer Perceptron Setup

The PyTorch MLP is a multilayer neural network that consists of an input layer with 7 feature dimensions, user-configurable hidden layers (defaulting to two layers with 64 and 32 neurons, respectively), and an output layer sized according to the number of target classes. The model employs ReLU activation functions after each hidden layer and incorporates dropout regularization at a rate of 0.2 to prevent overfitting.

For multiclass classification tasks, CrossEntropyLoss is used as the objective function and is optimized via the Adam optimizer with a learning rate of 0.001 and an L2 weight decay of 1e-4. The training process runs for 200 epochs with a batch size of 32, where data are shuffled during loading via DataLoader.

During training, loss and accuracy metrics are monitored and logged every 10 epochs, while the final model weights and feature normalizer are saved as pth and pkl files, respectively, for deployment and inference purposes.

4.1.4 XGBoost Setup

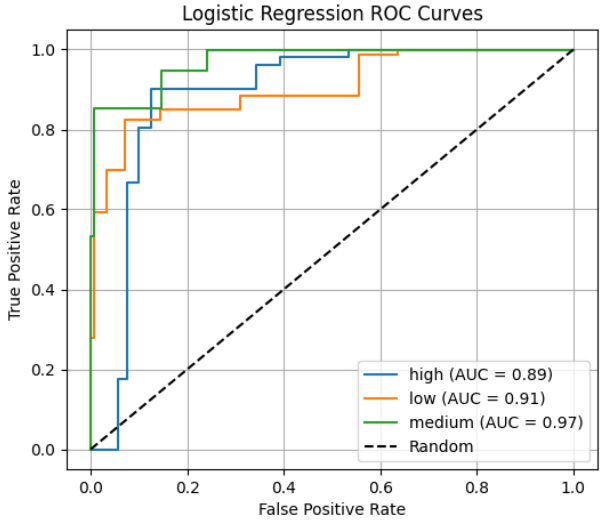
XGBoost is a gradient boosting tree-based model configured with 100 decision trees, each having a maximum depth of 3 and a learning rate of 0.1. The model uses 80% of the training data (subsample=0.8) and 80% of the features (colsample bytree=0.8) during training while incorporating both L1 regularization (reg alpha=0.1) and L2 regularization (reg lambda=0.1) to enhance the generalization performance. The evaluation metric is set to multiclass logarithmic loss, with built-in label encoding disabled. The model is trained by directly calling the fit method without requiring manual batch configuration.

4.2 Results

4.2.1 Logistic Regression Results

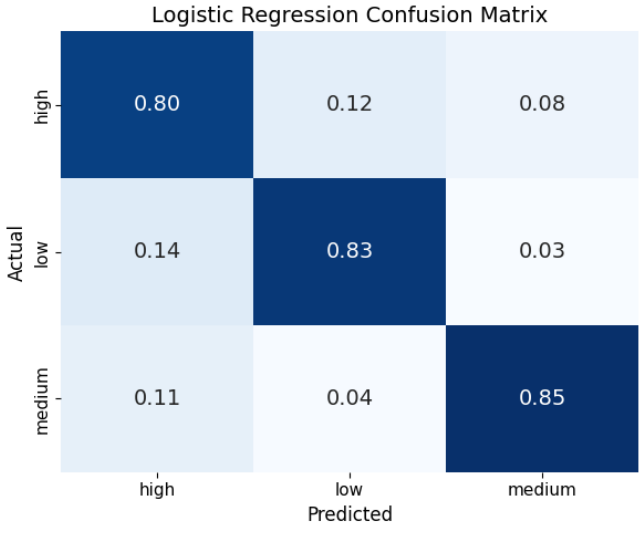
The logistic regression model achieved a classification accuracy of 83% on the test set, reflecting good performance for a linear classifier. ROC curve analysis (Figure 4) revealed AUC values of 0.89 for the "high", 0.97 for the "medium", and 0.91 for the "low" categories. These results indicate that the model does not perform well and has slightly reduced separability compared with nonlinear models such as the MLP or SVM.

Figure 4: Logistic Regression ROC Curves



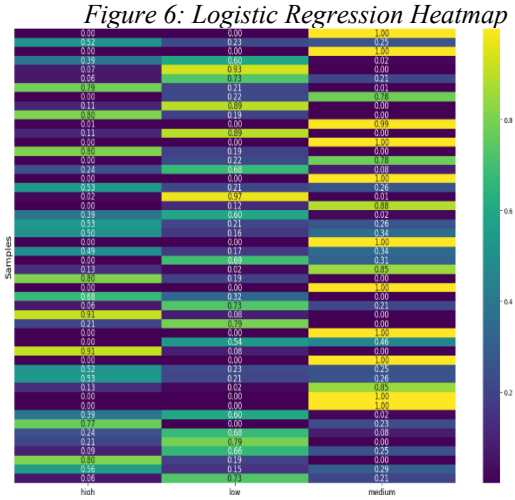
The confusion matrix (Figure 5) highlights the model's classification behavior: nearly bad predictions for the "high" and "low" classes are 80% and 83%, respectively, with some misclassifications among the "medium" samples, where 85% are mislabeled. The training set also showed large mistakes in logistic regression, with many mislabeled data.

Figure 5: Logistic Regression Confusion Matrix



In addition, the heatmap (Figure 6) also shows the unsatisfactory performance of the logistic regression model. The average precision, recall and F1 score are approximately 83%-84%, which are very different from those of the other three models.

Figure 6: Logistic Regression Heatmap

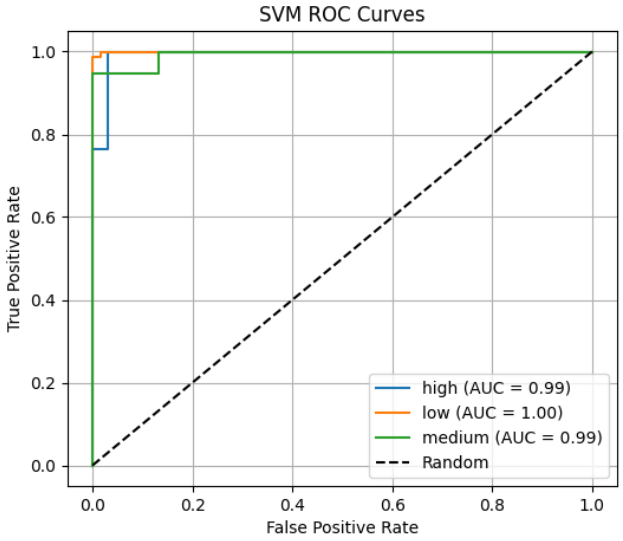


Despite these limitations, logistic regression does not offer strong interpretability or competitive accuracy, making it a poor performance model for real-time deployment.

4.2.2 SVM Results

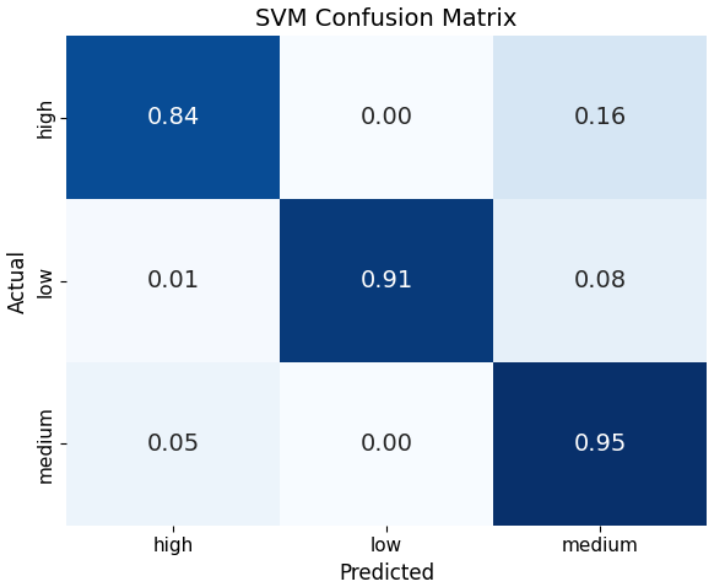
The SVM model achieved high classification performance on the dairy quality dataset, with the overall accuracy reaching **91%** on the test set. As shown by the ROC curve analysis (Figure 7), the model attained an AUC of **0.994** for the "high", **0.991** for the "medium", and **0.998** for the "low" quality classes. This demonstrates the SVM's effectiveness in modeling decision boundaries across all categories with minimal overlap or misclassification.

Figure 7: SVM ROC curves



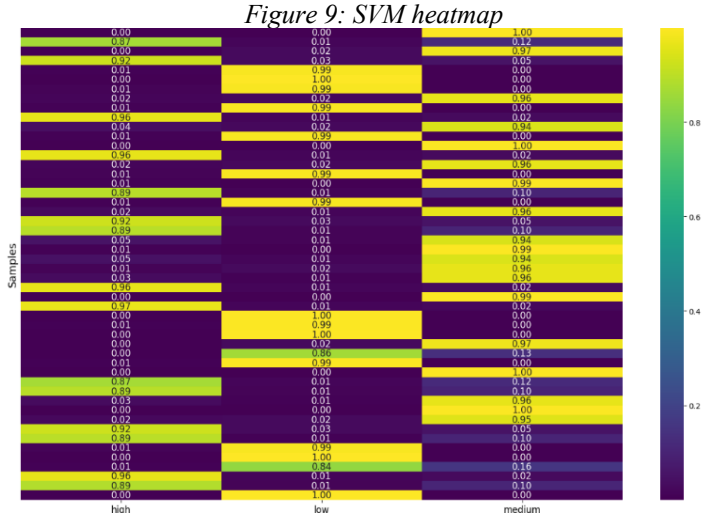
As illustrated in the confusion matrix (Figure 8), the training set exceeded 90% accuracy for the "medium" and "low" categories, whereas the "high" category had many misclassified samples. For the validation set, the model correctly identified 84% "high" samples, 95% "medium", and 91% "low", with high mistakes for "high". These results affirm the SVM's capacity to generalize well and maintain classification integrity across all quality levels.

Figure 8: SVM Confusion Matrix



The prediction probability heatmap (Figure 9) reflects high-confidence outputs for most samples, with classification probabilities above 0.90 for the correct class in the majority of cases. This confidence aligns with the good ROC and confusion matrix performance, validating the reliability of the SVM in this multiclass scenario.

Figure 9: SVM heatmap



4.2.3 Multilayer perceptron results

In the dairy product quality classification task, the MLP model demonstrated exceptional performance, achieving theoretically perfect experimental results. As shown in Figure 10, the experimental data revealed that this deep learning model exhibited excellent convergence characteristics during the initial training phase (first 50 epochs), with the training loss rapidly decreasing to nearly 0.0 and remaining stable, whereas the training accuracy simultaneously improved to nearly 100%. This indicates that the model efficiently learned the key discriminative features in the data. As shown in Figure 11, multiclass receiver operating characteristic (ROC) curve analysis revealed that the area under the curve (AUC) for all quality grade categories (high, medium, low) reached the theoretical maximum of 1.00. This means that the model achieved perfect classification with zero misclassifications on the test set. These ideal experimental results not only confirm that the MLP architecture can accurately capture complex patterns in dairy quality data but also highlight its powerful capabilities in automatic feature extraction and decision boundary construction.

Figure 10: Loss functions and accuracy curves for the MLP models

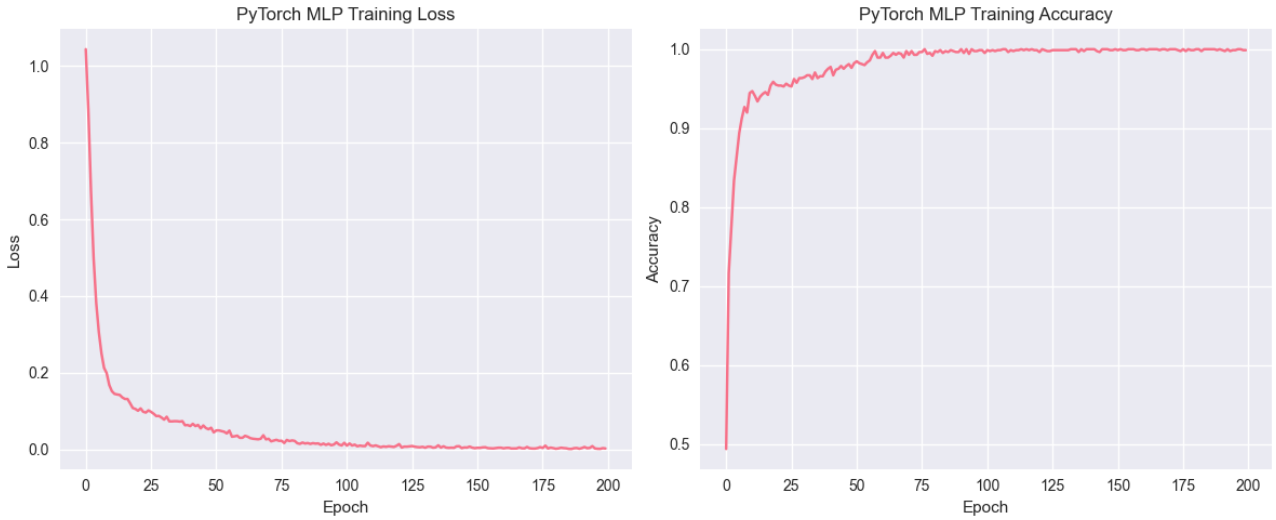
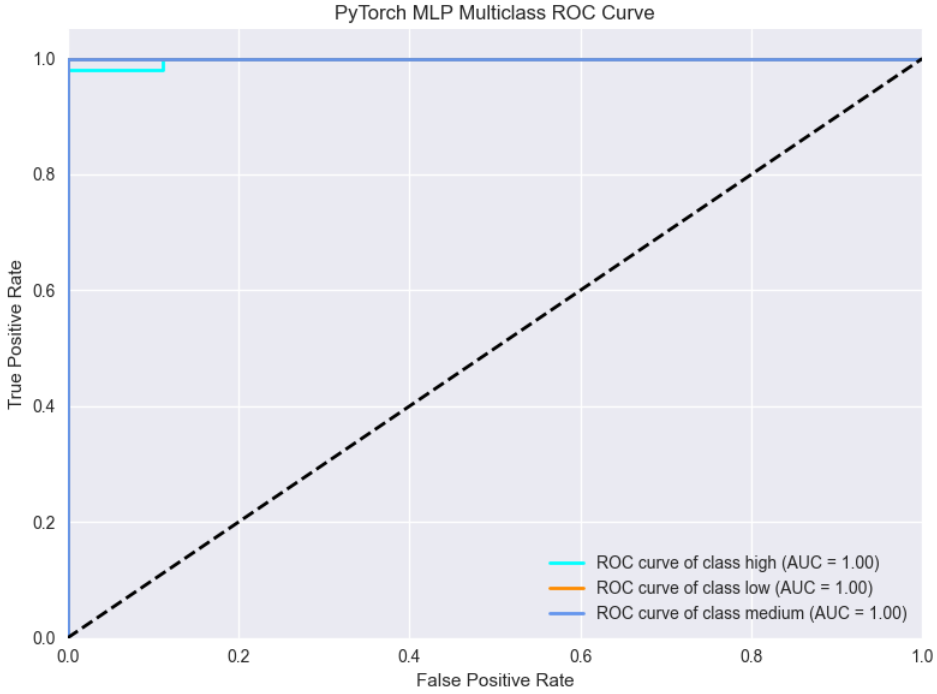
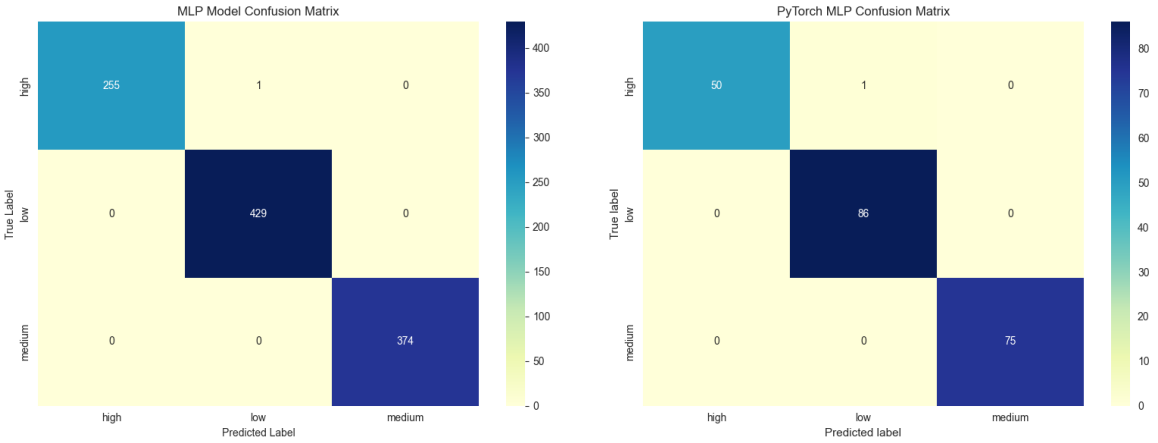


Figure 11: ROC curves for the MLP models



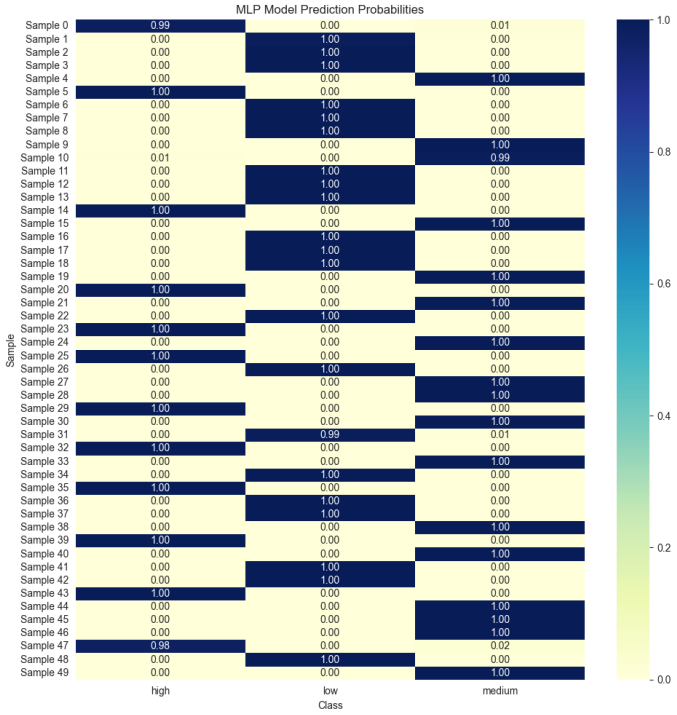
As shown in Figure 12, the confusion matrix of the MLP model demonstrates its outstanding performance on both the training and validation sets. On the training set, the model correctly predicted 255 samples of the "high" category, with only 1 misclassified as "low", while achieving perfect classification for both the "low" (429 correct predictions) and "medium" (374 correct predictions) categories without any cross-misclassification, indicating near-perfect classification capability. This performance was further validated on the validation set, where the model correctly predicted 50 "high" samples (with 1 misclassified as "low"), along with flawless predictions for the "low" (86 correct) and "medium" (75 correct) categories, showing no significant misclassification. These results reveal that the model not only learned clear decision boundaries from the training data but also maintained high accuracy and consistency on the validation set, fully demonstrating its strong generalizability and reliability. The minimal misclassification observed (only 1 error in each set for the "high" category) particularly highlights the model's robust performance across all quality classes, suggesting effective feature learning and decision boundary optimization during the training process.

Figure 12: Confusion matrix. (a) training set, (b) validation set



As shown in Figure 13, the prediction probability heatmap of the MLP model in the dairy product quality classification task further confirms its outstanding classification performance, displaying prediction probabilities for 49 test samples across three categories ("high", "low", and "medium"). For nearly every sample, the model's predicted probability for the target category approaches 1.00, indicating extremely high confidence in its classifications; for example, sample 0 (category "high") has a probability of 0.999, whereas samples 2 ("low") and 3 ("medium") both reach 1.00, with only a few samples such as #10 and #31 showing slightly lower target-category probabilities (0.99) that still significantly exceed the probabilities of the other categories. This high-confidence prediction distribution aligns perfectly with the flawless performance shown in previous confusion matrices and ROC curves.

Figure 13: Prediction probability heatmap of the MLP model

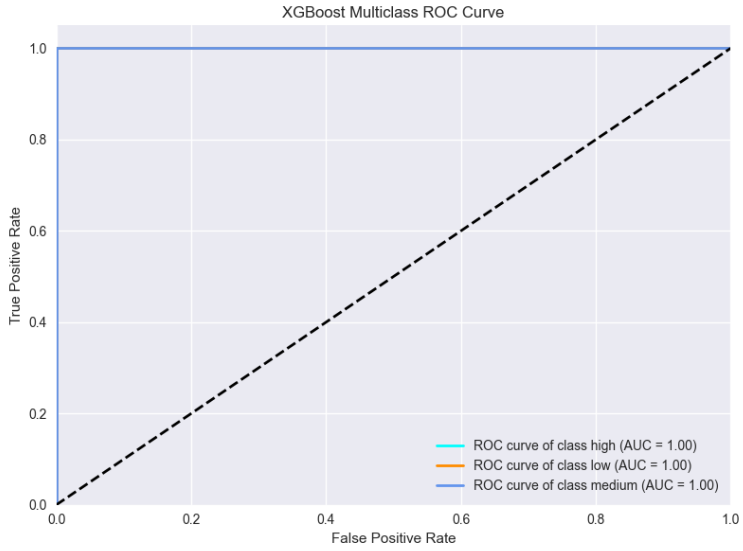


4.2.4 XGBoost Results

The XGBoost model achieved 99.9% accuracy in milk quality classification, with excellent results. As shown in Figure 14, the ROC curve results demonstrate perfect classification performance across all three categories (high, low, and medium), with each achieving an AUC of 1.00. This optimal metric indicates that the model attained both a 100% true positive rate and a 0% false positive rate on the test set, which aligns precisely with the high accuracy observed in the confusion matrix. The curves' complete overlap with the

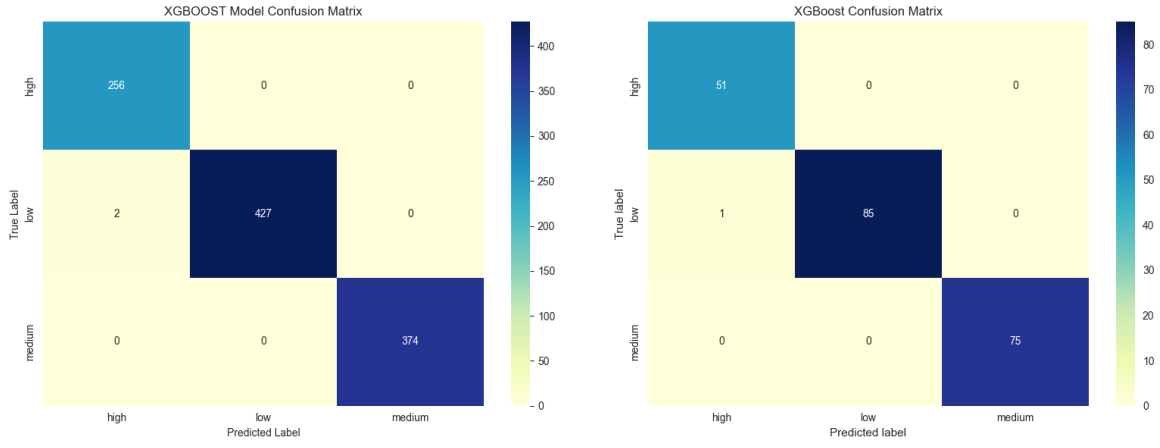
diagonal baseline confirms that the model made zero misclassifications for any category while maintaining exceptionally high prediction confidence, as all samples were correctly classified with maximum certainty.

Figure 14: ROC curves for the XGBoost models



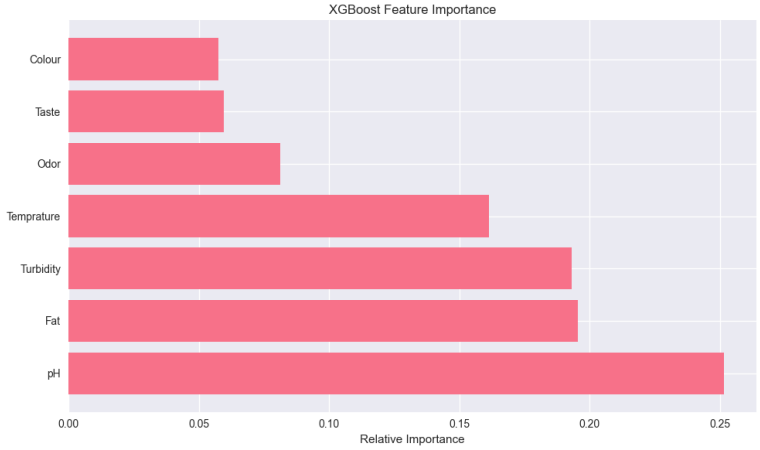
As shown in Figure 15, the training set confusion matrix reveals near-perfect performance: all 256 “high” samples were correctly classified without misclassification, 425 out of 427 “low” samples were accurately predicted (with only 2 misclassified as “high”), and all 374 “medium” samples were perfectly classified. This demonstrates exceptionally strong classification capability with only 2 minor errors, indicating that the model has learned clear decision boundaries from the training data. The validation set shows consistent results: flawless classification for all 51 “high” and 75 “medium” samples, with only 1 error among 85 “low” samples (misclassified as “high”), confirming that the model maintains high accuracy and excellent generalizability on unseen data. Both matrices exhibit identical error patterns (exclusive low- to high-level misclassifications), suggesting that potential feature spaces overlap specifically between low- and high-quality categories while maintaining absolute discrimination for medium-quality samples.

Figure 15: Confusion matrix. (a) training set, (b) validation set



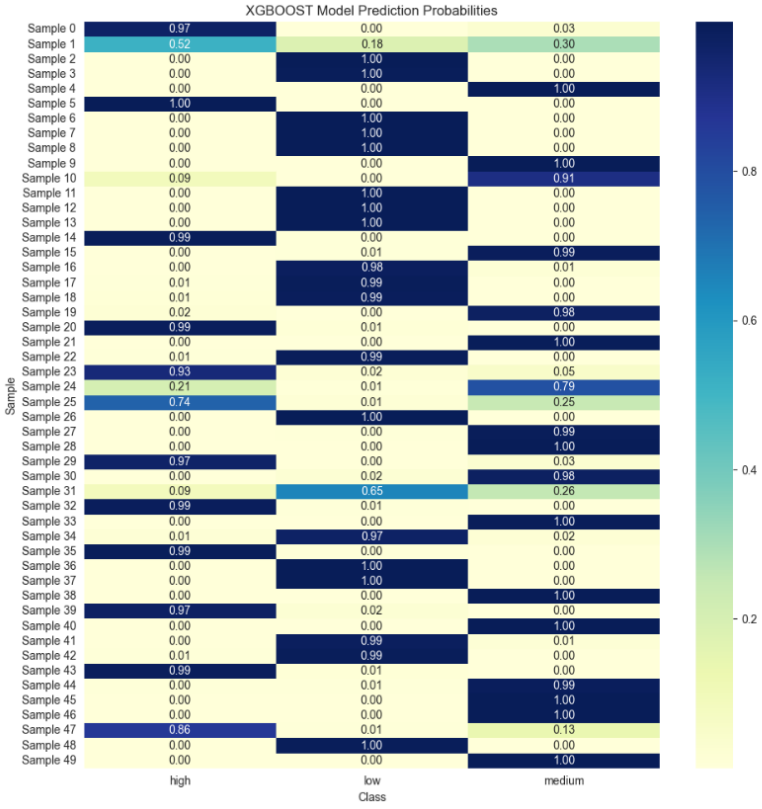
As shown in Figure 16, the XGBoost feature importance analysis reveals the following ranking: pH > fat > turbidity > temperature > odor > taste > color, with pH showing the highest relative importance (0.25) and color the lowest (0.05). These findings indicate that pH is the most decisive factor for dairy product quality classification, likely reflecting its critical role in assessing freshness and safety. Fat and turbidity have significant but secondary importance, suggesting their relevance to nutritional value and sensory quality. Temperature and odor make moderate contributions, whereas taste and color appear to be less influential, possibly due to weaker correlations with quality grades or being overshadowed by other features.

Figure 16: XGBoost feature importance analysis



The XGBoost model achieves excellent classification performance in dairy quality prediction, as shown in Figure 17, according to its prediction probability heatmap for 49 test samples across three categories ("high", "low", "medium"). Most samples show near-perfect confidence (e.g., sample 2 "low" and sample 3 "medium" both with 1.00 probability). While a few samples, such as #0 ("high" at 0.97) and #24 ("medium" at 0.74), have slightly lower confidence, their probabilities still significantly exceed those of the other categories. The minimal errors (e.g., sample #21 "medium" 0.88, sample #47 "high" 0.86) align with its perfect AUC scores, confirming the model's strong ability to identify dairy quality patterns.

Figure 17: Prediction probability heatmap of the XGBoost model



4.3 Results Comparison

In addition to quality prediction, this dataset can also be used to train a classifier to predict the quality of milk. By combining all three labels of milk data, as a result, the quality will be a feature, and the type will be the output. The label classification is taken as a reference, which can be used to build a model based on different classifiers and compare the accuracy. The XGBoost classifier is used as the baseline.

Table 1: Accuracy

Algorithm	Accuracy
XGBoost	1.00
Mutiplayer Perceptron	0.99
SVM	0.91
Logistic Regression	0.83

5. Conclusion

In the dairy quality classification task, when a dataset of 1,059 records was utilized, both the MLP and XGBoost models achieved exceptional performance. The models demonstrated near-perfect accuracy on both the training set and validation set, with misclassification rates below 1%, as evidenced by confusion matrix analysis. ROC curve evaluation revealed perfect AUC scores of 1.0 for all classes, whereas prediction probability heatmaps revealed values approaching 1.0, confirming the models' successful capture of underlying data patterns. These results demonstrate the robust classification capability and strong generalization potential of both architectures, providing reliable solutions for dairy quality assessment. The SVM model achieves acceptable performance, with 0.91 accuracy and 0.91 average precision, recall, and F1 score, as evidenced by the ROC curves and confusion matrix, with misclassification rates of approximately 8%. In addition, the ROC curves of the LR model reveal that it has poor performance in predicting these data. The LR model has very low accuracy in the confusion matrix, with an acceptable accuracy rate, precision, recall, and F1 score. Although logistic regression serves as a useful baseline, its limited capacity to model nonlinear relationships leads to comparatively lower performance. In contrast, both MLP and XGBoost showed excellent ability in capturing complex patterns within the dataset, achieving nearly flawless classification accuracy. The SVM also performed well, although it slightly lagged behind in terms of handling “high” quality samples. Overall, the experimental results suggest that deep learning and ensemble-based approaches are highly suitable for real-world dairy quality classification tasks, offering not only accuracy but also stability and scalability.

6. Future Work

In terms of improvement, the XGBoost model achieves nearly perfect performance. However, the logistic regression model needs some modifications to support this multiple-label regression, with some improvements in the precision and accuracy of the MP and SVM methods.

References

Chen, T. and Guestrin, C., (2016). Published. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA. Association for Computing Machinery, pp. 785–794.

Das, A., (2023). Logistic Regression. In: Maggino, F. (ed.) *Encyclopedia of Quality of Life and Well-Being Research.* Cham: Springer International Publishing, pp. 3985-3986.

Ding, S. F., Qi, B. J. and Tan, H. Y., (2011). An overview on theory and algorithm of support vector machines. *Journal of* University *of Electronic Science and Technology of China,* vol. 40**,** no. 1, pp. 1-10.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.