Current Research in Food Science 12 (2026) 101353

Contents lists available at ScienceDirect

Current Research in Food Science

journal homepage: www.sciencedirect.com/journal/current-research-in-food-science

Predicting aflatoxin M_1 in raw milk using machine learning and basic measurements

Haohan Ding [a,c,*], Long Wang [a], Xiaodong Song [b], Xiaohui Cui [a,c,d], David I. Wilson [d,e], Wei Yu [f], Cheng Zhang [g], Guanjun Dong [b]

[a] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, China
[b] State Administration for Market Regulation, Huhehaote, 011517, China
[c] Science Center for Future Foods, Jiangnan University, Wuxi, 214122, China
[d] School of Cyber Science and Engineering, Wuhan University, Wuhan, 430072, China
[e] Department of Data Science and AI, Auckland University of Technology, Auckland, 1010, New Zealand
[f] Department of Chemical & Materials Engineering, University of Auckland, Auckland, 1010, New Zealand
[g] School of Environment & Ecology, Jiangnan University, Wuxi, 214122, China

ARTICLE INFO

Handling Editor: Dr. Maria Corradini

Keywords:
Aflatoxin M_1
Machine learning
Prediction model
Raw milk

ABSTRACT

Aflatoxin M_1 (AFM_1) is a carcinogenic and teratogenic mycotoxin that may be present in raw milk. Therefore, continuous monitoring of AFM_1 levels is essential to ensure dairy safety and regulatory compliance. Although laboratory-based analytical techniques such as ELISA and LC-MS/MS offer high accuracy, their cost, sample preparation requirements, and dependence on specialized personnel make them less practical for high-frequency or large-volume screening in dairy processing facilities. This creates a need for complementary, cost-effective prescreening approaches. This study proposed a qualitative AFM_1 prediction method based on routinely measured physicochemical indicators of raw milk, combined with machine learning algorithms. Five classical machine learning models were evaluated under a binary classification framework to determine whether AFM_1 levels exceed the regulatory threshold. Experimental results show that the multilayer perceptron achieves an accuracy and negative-sample recall rate above 80%, demonstrating the potential of machine learning as an effective prescreening tool for AFM_1. The findings provide a feasible direction for supporting rapid, economical, and large-scale monitoring of raw milk safety.

1. Introduction

Milk is widely consumed worldwide due to its nutritional value and health benefits (Picciano, 2001). To meet the increasing demand, dairy factories collect raw milk from multiple resources, making it essential to assess its physicochemical properties and ensure compliance with safety standards. Among various contaminants of concern, aflatoxins have received particular attention due to their severe health implications.

Aflatoxin B_1 (AFB_1), produced primarily by *Aspergillus flavus* and *Aspergillus parasiticus* (Zheng et al., 2013), is commonly found in animal feed (Rodrigues and Naehrer, 2012). When ingested by dairy cows, AFB_1 is metabolized into AFM_1, which is subsequently excreted into milk (Tadesse et al., 2020; Xiong et al., 2020). Both toxins are classified as carcinogenic, mutagenic, and teratogenic (Smoke and Smoking, 2004). To protect consumer health, regulatory agencies have established strict limits for AFM_1 in milk, such as 0.05 μg/kg in the European Union and Codex Alimentarius and 0.5 μg/kg in the United States (Omar, 2016). Ensuring that raw milk meets these limits is therefore of great importance.

A range of analytical methods, including enzyme-linked immunosorbent assay (ELISA), thin-layer chromatography (TLC), immunoaffinity column extraction, high-performance liquid chromatography with fluorescence detection (HPLC-FLD), and liquid chromatography–tandem mass spectrometry (LC-MS/MS), are currently used for AFM_1 determination (Singh et al., 2022). While these techniques provide high sensitivity and accuracy, they often require specialized equipment, trained personnel, and extensive sample preparation. As a result, applying these methods to the large number of samples processed raw milk in dairy factories can be cost-intensive and time-consuming, making them less suited for high-throughput or continuous monitoring

* Corresponding author. School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, China.
E-mail address: dinghaohan@jiangnan.edu.cn (H. Ding).

https://doi.org/10.1016/j.crfs.2026.101353
Received 4 July 2025; Received in revised form 13 February 2026; Accepted 13 February 2026
Available online 16 February 2026

Table 1
Examples of negative and positive samples for AFM1 After Treatment: An Analysis of Key Milk Components (Total Colony Count (TCC), Protein, Fat, Lactose, Total Solids (TS), Freezing Point (FP), Somatic Cells (SC), Non-Fat Milk Solids (N-FMS), Acidity, Psychrophilic Bacteria (PB), Viscosity, Relative Density (RD), Milk Temperature (T).

AFM_1	TCC	Protein	Fat	Lactose	TS	N-FMS	Acidity	FP	SC	PB	Viscosity	RD	T
	cfu/mL	g/100g	g/100g	g/100g	%	%	°T	°C	cells/mL	cfu/mL	mPa·s	g/cm^3	°C
1	2.1	3.35	3.91	4.97	12.65	8.73	12.89	−0.520	23.2	2200	682	1.030	4.2
1	2.2	3.18	3.81	4.68	12.20	8.33	13.07	−0.522	49.0	1000	1158	1.029	3.5
…	…	…	…	…	…	…	…	…	…	…	…	…	…
0	1.8	3.33	4.81	4.96	13.64	8.82	13.35	−0.535	26.5	800	1103	1.030	3.2
0	2.2	3.47	4.17	4.99	13.17	9.02	13.56	−0.545	20.2	1000	1103	1.031	3.6

applications.

With the advancement of artificial intelligence, machine learning (ML) has shown strong capabilities in pattern recognition, nonlinear modeling, and data-driven prediction across many food safety applications. ML has been successfully integrated with electronic sensors (Wu et al., 2019), spectral analysis (Mansuri et al., 2022), and imaging systems (Jiang et al., 2021), achieving high accuracy in detecting food contaminants. Prior research includes the use of ML with infrared spectroscopy for detecting residues in milk powder (de Freitas et al., 2021), as well as electronic tongue systems for differentiating food quality (Wang et al., 2019). Additionally, environmental ML models have demonstrated high predictive performance for regional mycotoxin contamination in crops (Castano-Duque et al., 2022).

In dairy processing, routine measurements such as fat, protein, lactose, and acidity are already collected at low cost. Using these easily available physicochemical indicators to infer AFM_1 contamination offers the potential for a practical prescreening tool that reduces the reliance on expensive confirmatory testing. Motivated by this, the present study investigates whether the composition of raw milk contains sufficient information to qualitatively predict AFM_1 compliance status. Using routinely measured milk constituents as input features, we develop a machine learning framework to classify whether AFM_1 levels exceed the regulatory threshold. This approach aims to provide dairy factories with a rapid and economical screening method that can prioritize samples for further laboratory analysis.

2. Materials and methods

2.1. Raw milk samples

A dairy factory collected nearly one million raw milk measurement records between January 1, 2022 and October 31, 2024. After missing entries and outliers were removed, more than 40,000 representative samples were retained for analysis, from which a final data set was constructed. The dataset includes sixteen numerical features: total colony count, protein, fat, lactose, total solids, non-fat milk solids, acidity, freezing point, somatic cells, psychrophilic bacteria, viscosity, relative density, milk temperature, country, province and AFM_1 concentration. Examples of these features are shown in Table 1.

Following the EU regulatory threshold, AFM_1 concentrations 0.05 μg/L were labeled as unqualified (assigned label 1), while values $<$ 0.05 μg/L were labeled as qualified (label 0). Although this appears counterintuitive (where positive typically indicates unqualified), we retained this mapping to remain consistent with factory internal records. To avoid confusion, throughout this paper we use the terminology unqualified samples (label 1) and qualified samples (label 0). The final dataset contained nearly 500 unqualified samples.

2.2. Data measurement method and meaning

This section outlines the methods of measurement and the meaning of each feature in the raw milk samples used in this study.

(1) The study used the enzyme-linked immunosorbent assay (ELISA) method as specified in the Chinese national food safety standard (GB/T 5009.24-2016) for detecting AFM_1 in raw milk. The colorimetric signal of the conventional enzyme-linked immunosor-bent assay (ELISA) is produced by horseradish peroxidase (HRP) or alkaline phosphatase (ALP) by catalyzing colorless organic dyes to produce colored chemical compounds (Bao et al., 2021).
(2) A near-infrared spectrometer was used to measure protein, fat and lactose in raw milk (Šašić and Ozaki, 2001). Protein and lactose content are critical for assessing milk's nutritional value, while fat in the form of tiny globules affects the taste and flavor of milk.
(3) The total colony count quantifies bacteria in milk. The sample is treated with enzymes and centrifuged to remove protein and fat, leaving only the bacterial precipitate. Bacteria are stained with a fluorescent dye that binds to nucleic acids, emitting a green signal. A flow cytometer distinguishes bacteria from other particles based on laser scattering and fluorescence for quantitative analysis (McKinnon, 2018).
(4) Total solids refer to the mass of the remaining material after drying a 100g sample of milk at 102 – 105 °C, including protein, fat, lactose and minerals. Non-fat milk solids are the total solids that exclude fat and reflect the nutritional quality of milk (Clark et al., 1989).
(5) The acidity of raw milk includes both natural and fermentation acidity. Natural acidity arises from components such as proteins and phosphates, while fermentation acidity increases during storage because of microbial activity. We measure acidity using the potentiometric titer method, according to GB 5009.239-2016, Acidity indicates freshness and quality of milk, and increased acidity suggests contamination or improper storage (Lu et al., 2013).
(6) The freezing point of milk indicates whether water has been added. The addition of water reduces the freezing point, deviating from the normal range and signaling potential adulteration (Harding, 1995).
(7) The somatic cell count is indirectly predicted by spectral analysis in the near-infrared range (1100 – 2500 nm), providing insight into milk quality (Tsenkova et al., 2001).
(8) Psychrophilic bacteria thrive in cold temperatures (below 7 °C) and can decompose milk components, producing undesirable odors and accelerating spoilage (Oliveira et al., 2015). These bacteria are counted using a flow cytometer (McKinnon, 2018).
(9) Viscosity during fermentation is measured at different coagulation temperatures (30 °C and 40 °C) and protein concentrations (3.4 % and 5.1 %). This feature reflects the degree and quality of milk fermentation.
(10) Relative density is the ratio of milk's mass to the mass of an equal volume of water at a specific temperature. It helps identify adulteration and assess milk quality.

(11) The temperature of raw milk on arrival at the factory is measured using a thermometer, which is important for assessing the conditions of milk storage.
(12) Country and province which the collection area belongs identify geopolitical differences in data collection.

2.3. Preliminary exploratory analysis

Principal Component Analysis (PCA) was first conducted on thirteen physicochemical features (excluding AFM_1) to explore their dominant directions of variance. It is important to note that PCA was not used for dimensionality reduction in subsequent predictive modeling; instead, it served only as an exploratory tool to assess characteristic relationships, evaluate variance contributions, and visualize sample clustering patterns (Greenacre et al., 2022).

Cumulative variance ratios and loading plots were used to identify features that contribute strongly to the main components. In addition, unqualified samples were grouped by milk source to examine whether certain geographic origins presented higher contamination rates, which may be related to regional exposure to aflatoxin related to feed.

2.4. Model selection and hyperparameter tuning

In this study, we used five different models—Linear Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost) (Chen, 2016), and Multi-Layer Perceptron (MLP)—to predict the levels of AFM_1 in raw milk.

As illustrated in Fig. 8, in the Multi-Layer Perceptron (MLP) architecture, we employ the Rectified Linear Unit (ReLU) activation function following the hidden layers (Nair and Hinton, 2010). ReLU is a widely adopted nonlinear activation function, which operates by outputting zero when the input is less than or equal to zero and directly passing positive inputs unchanged. This function facilitates the introduction of nonlinear transformations, enabling the neural network to capture more complex patterns and features. In the output layer, we apply the Softmax function (Bridle, 1989), which normalizes the exponential values, ensuring that the model's outputs are constrained between 0 and 1, thereby representing the probability of the positive class. For the optimization process, we use Binary Cross Entropy (BCE) loss, a standard loss function for binary classification tasks (Mao et al., 2023). BCE measures the discrepancy between the predicted probabilities and the actual labels, guiding the model to minimize the error and improve its classification accuracy.

These models represent a broad spectrum of approaches that include classical statistical methods, traditional machine learning algorithms, and ensemble learning methods. This diversity allows for a comprehensive evaluation of the performance of different methods in the task of AFM_1, providing a comparative basis for model optimization.

In response to the importance of geographical distribution highlighted in previous studies (Bilandžić et al., 2022; Hernández-Martínez and Navarro-Blasco, 2015; Mudannayake et al., 2024; Peña-Rodas et al., 2018), the occurrence of AFM_1 in raw milk is primarily driven by the presence of AFB_1 in feed. The initial contamination level of AFB_1 in grains is influenced by climatic conditions (temperature and humidity), as well as harvesting and storage practices. After ingestion, AFB_1 is metabolized in the liver of dairy cows to AFM_1, and a defined carry-over rate governs the transfer from feed to milk, with individual variation among animals. The amount of AFM_1 excreted into milk may also depend on milk yield and metabolic status. Therefore, the geographic origin of each milk sample was incorporated as an additional predictor in the model. The raw location information, consisting of "province" and "city", was treated as a categorical variable. Since machine learning models cannot directly process textual categorical attributes, the geographical information was encoded prior to training. Specifically, we applied target encoding (Prokhorenkova et al., 2018) for tree-based models (RF and XGBoost) and one-hot encoding (LeCun et al., 2015) for linear and neural network models (LR, SVM, MLP). This dual strategy prevents high-dimensional sparse encoding in ensemble models and ensures numerical stability in models sensitive to feature scaling.

Table 2
List of primary hyperparameters and their corresponding search spaces for the machine learning models.

Model	Hyperparameter	Search Space
LR	*C* (Inverse regularization)	{0.2, 0.5, 1, 2, 5}
	Penalty type	L2
	Solver	liblinear
RF	Number of estimators	{50, 100, 150, 200, 300}
	Maximum depth	{0, 10, 20, 30, 40, 50}
	Min. samples for split	{2, 5, 10, 15}
	Min. samples for leaf	{1, 2, 4, 6}
	Max. features	{auto, sqrt, log2, None}
SVM	*C* (Inverse regularization)	{0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5}
	γ (Kernel coefficient)	{1, 0.7, 0.5, 0.45, 0.4, 0.35, 0.3, 0.2, 0.1}
	Kernel type	RBF
XGBoost	Number of estimators	{100, 200, 300}
	Maximum depth	{9, 11, 13, 15}
	Learning rate	{0.01, 0.035, 0.05}
	Subsample ratio	{0.6, 0.8}
	Column subsample ratio	{0.6, 0.8}
	γ (Min. loss reduction)	{0.3, 0.4}
MLP	Number of layers	{1, 2, 3, 4}
	Hidden units	{32, 64, 128, 256}
	Dropout rate	{0.1, 0.2, 0.25, 0.3}
	Learning rate	$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$
	Batch size	{16, 32, 64, 128}
	Threshold	{0.3, 0.4, 0.5, 0.6, 0.7}

To ensure a fair and robust comparison, the hyperparameters of all models were tuned systematically. The MLP model was optimized using the Tree-structured Parzen Estimator (TPE) sampler within the Optuna framework (Bergstra et al., 2011), allowing efficient search over a complex space. The remaining models (SVM, RF, XGBoost, LR) were tuned using standard Grid Search with cross-validation (Soper, 2021). The detailed search spaces and parameter ranges for each model are presented in Table 2.

2.5. Model evaluation

Given the significant imbalance between the positive and negative samples in our dataset, we took steps to mitigate this issue during model training. We used under-sampling of the positive samples (Yen and Lee, 2009), ensuring that the number of positive samples matched the number of negative samples, to balance the dataset. This was done by randomly selecting an equal number of positive samples to match the negative samples and combining them for training. This approach addresses the potential performance degradation caused by the class imbalance.

The dataset was divided into three sets for training, validation and testing, following the traditional 70 % training, 15 % validation, and 15 % test ratio. The training set (70%) was used for fitting model parameters. The validation set (15%) was used exclusively during hyperparameter optimization and early stopping. The test set (15%) was held out entirely during training and tuning. It was only used once after selecting the optimal hyperparameters to report the final performance in the Results section. To optimize the performance of the model, we employ a hyperparameter search strategy (Liashchynskyi and Liashchynskyi, 2019), in which we tested various combinations of hyperparameters to identify the optimal configuration for each model.

Before model training, all numerical features (except AFM_1 labels) were standardized using z-score normalization to ensure that no single feature dominates the objective function due to scale differences. This step is particularly critical for algorithms sensitive to feature magnitudes, such as SVM and LR. The normalization is defined as:

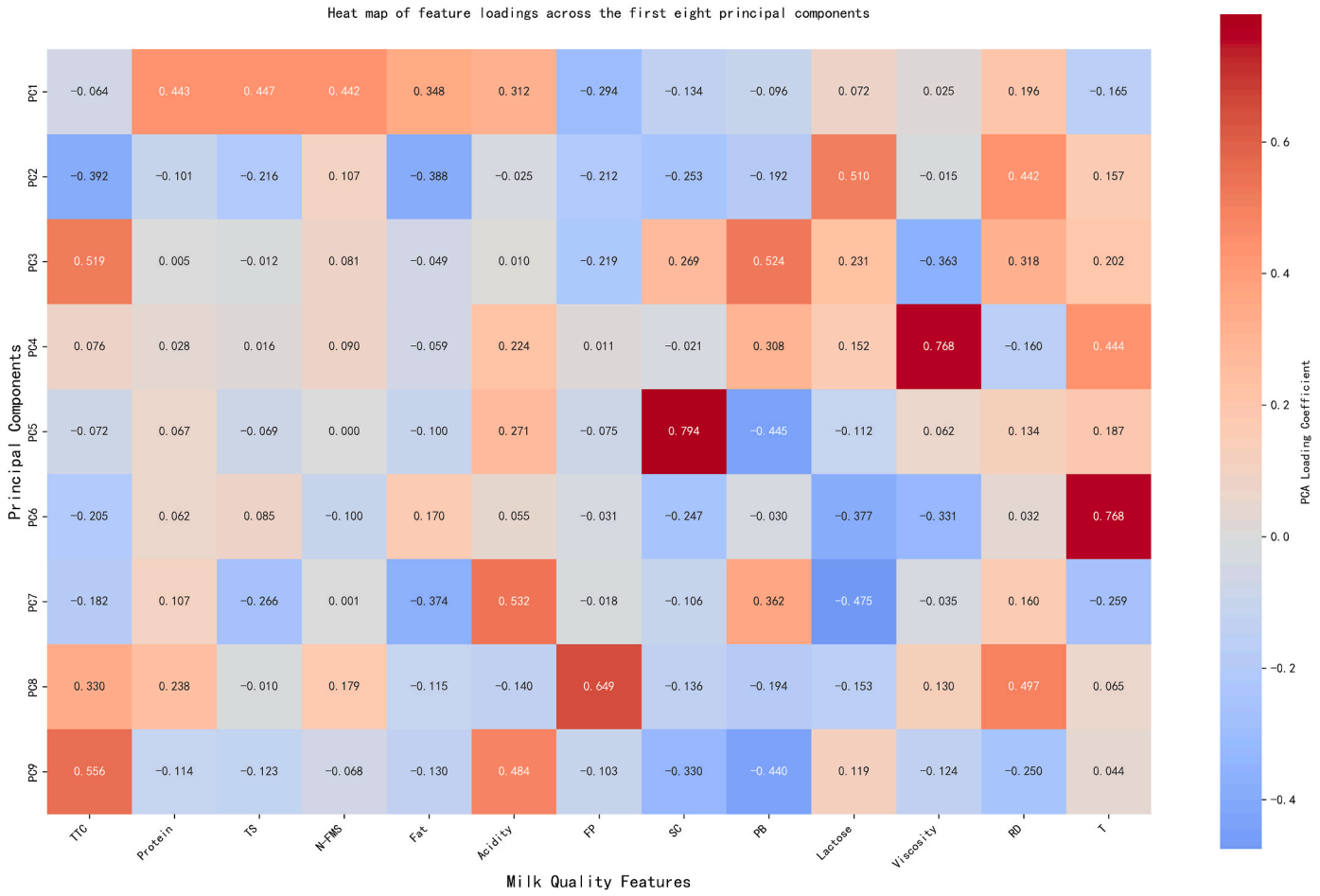


Fig. 1. Heat map of feature loadings across the first eight principal components.

$$\overline{x} = \frac{x - \mu}{\sigma}$$

where x is the original feature value, μ is the mean and σ is the standard deviation. Outliers were removed using the factory's internal quality control thresholds, and samples with missing AFM_1 labels were discarded entirely.

To further assess the generalizability of the model beyond the main dataset, an external independent dataset was used for additional validation. This dataset was collected by a dairy company in 2018, consisting of 137 AFM_1 negative and 8634 AFM_1 positive samples. After identifying the best-performing model through 100 repeated experiments on the main dataset, the optimal model (with fixed hyperparameters and architecture) was retrained on the entire training set and subsequently evaluated on the 2018 dataset. This external validation step enables assessment of temporal robustness and geographic transferability, addressing potential dataset-shift issues and providing stronger support for real-world applicability.

In order to reduce accidental errors in the experiment, we conducted 100 experiments on the optimal model obtained by a hyperparameter search, randomly selected 100 different data sets consisting of negative samples and positive samples for model evaluation, and calculated the variance to measure the stability of the model.

For performance evaluation, we used standard binary classification metrics (Hossin and Sulaiman, 2015):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

True Positives (TP) refer to the number of positive samples correctly predicted as positive, while True Negatives (TN) represent the number of negative samples correctly predicted as negative. False Positives (FP) are the negative samples incorrectly predicted as positive and False Negatives (FN) are the positive samples incorrectly predicted as negative.

2.6. Interpretability

To overcome the "black-box" nature of complex models like MLP and XGBoost (Ribeiro et al., 2016), we utilized SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) values to interpret model predictions. SHAP assigns each feature an importance value for a particular prediction based on cooperative game theory. For every predicted sample, the model generates a prediction value, and the SHAP value represents the contribution of each feature to that prediction.

Assume that the i th sample is X_i, the j feature of the i sample is X_{ij}, the model's predicted value for the sample is y_i, and the mean of the target variables of all samples in the entire model is y_{base}, then the SHAP value satisfies the following equation:

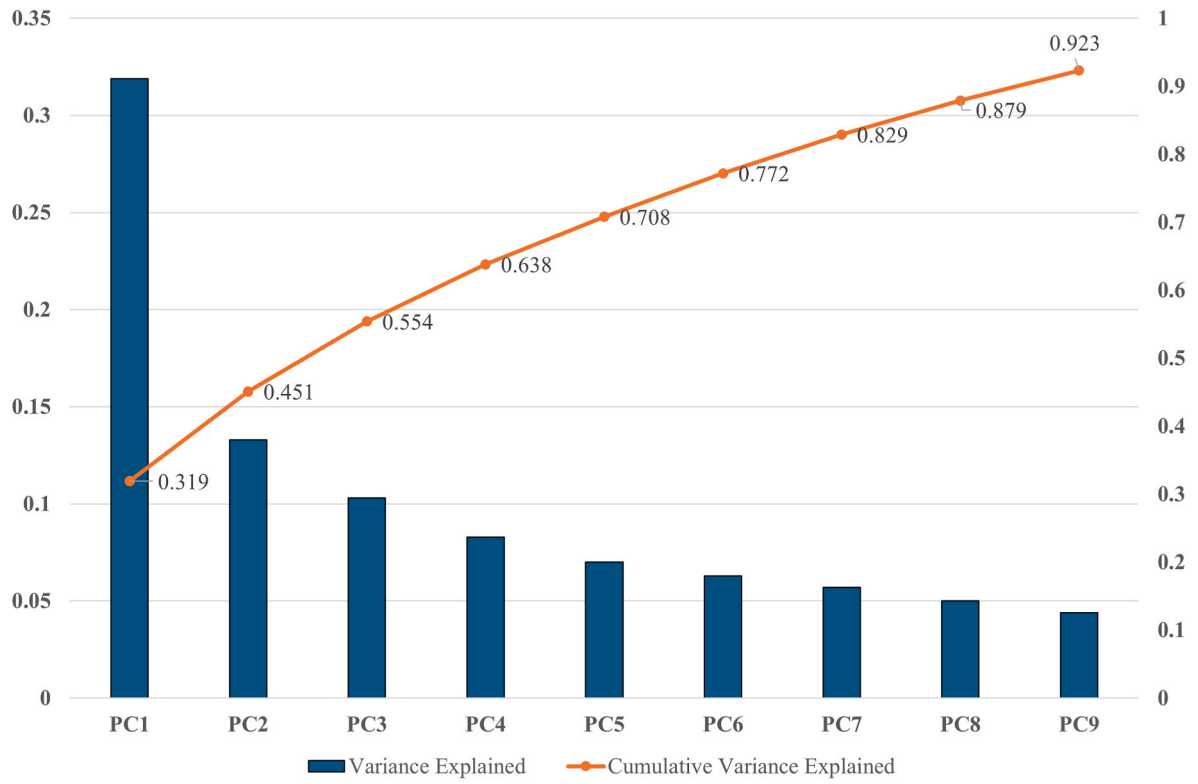


Fig. 2. Variance explained by the first nine principal components derived from standardized milk quality variables. Bars represent individual variance explained; orange line indicates cumulative variance.

$$y_i = y_{base} + f(X_{i1}) + f(X_{i2}) + \ldots + f(X_{ik}) \quad (1)$$

Where $f(X_{ij})$ is the SHAP value of X_{ij}. Intuitively, $f(X_{i1})$ is the contribution of the first feature in the i sample to the final predicted value y_i. When $f(X_{j1}) > 0$, it means that the feature improves the predicted value and plays a positive role. Conversely, when $f(X_{ij}) < 0$, it means that the feature reduces the predicted value and has a negative effect.

To comprehensively assess the consistency and stability of feature contributions, SHAP values were computed using the model's validation results on the additional external dataset. This analysis enables evaluation of whether the feature importance patterns learned during model training remain stable when the model is applied to data from different geographical origins. By comparing SHAP distributions between the internal validation set and the external 2018 dataset, we examined the robustness of key predictors, identified potential shifts in feature influence, and assessed the model's capacity to generalize its decision logic beyond the training environment. This cross-dataset interpretability evaluation provides stronger evidence that the model's predictions are not merely dataset-specific artifacts but instead reflect biologically and operationally meaningful patterns in raw milk quality attributes.

3. Results and discussion

3.1. Analysis of AFM_1 and feature variables

In this study, AFM_1 concentrations were converted into binary labels using a threshold of 0.05 μg/kg, where AFM_1 positive samples (< 0.05 μg/kg) were encoded as 0, and AFM_1 negative samples (≥ 0.05 μg/kg) were encoded as 1. This encoding is applied consistently in PCA, model evaluation, and SHAP analyzes.

To investigate the relationships between AFM_1 and raw milk physicochemical indicators, all negative samples were used as a baseline and positive samples were randomly resampled to mitigate class imbalance. The PCA was then performed on the full dataset. The principal component loading heat map (Fig. 1) and the cumulative variance contribution (Fig. 2) were obtained.

As shown in Fig. 2, no principal component explains a dominant portion of the variance, indicating that conventional compositional features (protein, fat, lactose, etc.) are not strongly numerically correlated with AFM_1. This is consistent with previous findings that report that AFM_1 contaminated feed did not have significant effects on the fat, protein, or lactose content of milk (Xiong et al., 2020).

PC1 exhibits high loadings on protein, total solids, non-fat milk solids, fat, and acidity (0.44, 0.45, 0.44, 0.35, and 0.31), reflecting the core nutritional attributes. PC2 captures lactose (0.51) and milk temperature (0.44) positively and total colony count (−0.39) negatively, indicating relationships with microbial quality and processing temperature. PC3 and PC5 mainly represent indicators of microbial contamination, such as colony count (0.52) and psychrophilic bacteria (0.52) for PC3, and somatic cell count (0.79) for PC5. PC4 (viscosity, 0.79) appears to be associated with fermentation-related bacterial activity, while PC6 (milk temperature, 0.77) suggests effects of external factors such as handling and storage. PC8 shows relevant contributions from relative density (0.50) and freezing point (0.65).

In general, PC1 and PC2 capture nutritional characteristics (45 % cumulative variance), PC3–PC5 reflect microbial contamination and PC6–PC9 reflect external environmental factors. However, none clearly explain the variations in AFM_1, likely due to the heterogeneous origins of the dataset. Milk composition varies by region, breed and production practices, potentially masking subtle effects related to AFM_1.

To reduce regional confounding, we further analyzed three regions with high sample counts and substantial negative-sample proportions. As shown in Fig. 3a–c, Region A had 18 negative samples among 407 samples, Region B had 25 among 598, and Region C had 23 among 523. Protein, fat, and acidity were compared using mean and percentile statistics (Fig. 3). The region A negative samples generally showed lower protein and fat, whereas Regions B and C showed higher protein and fat distributions. In all regions, negative AFM_1 exhibited consistently lower

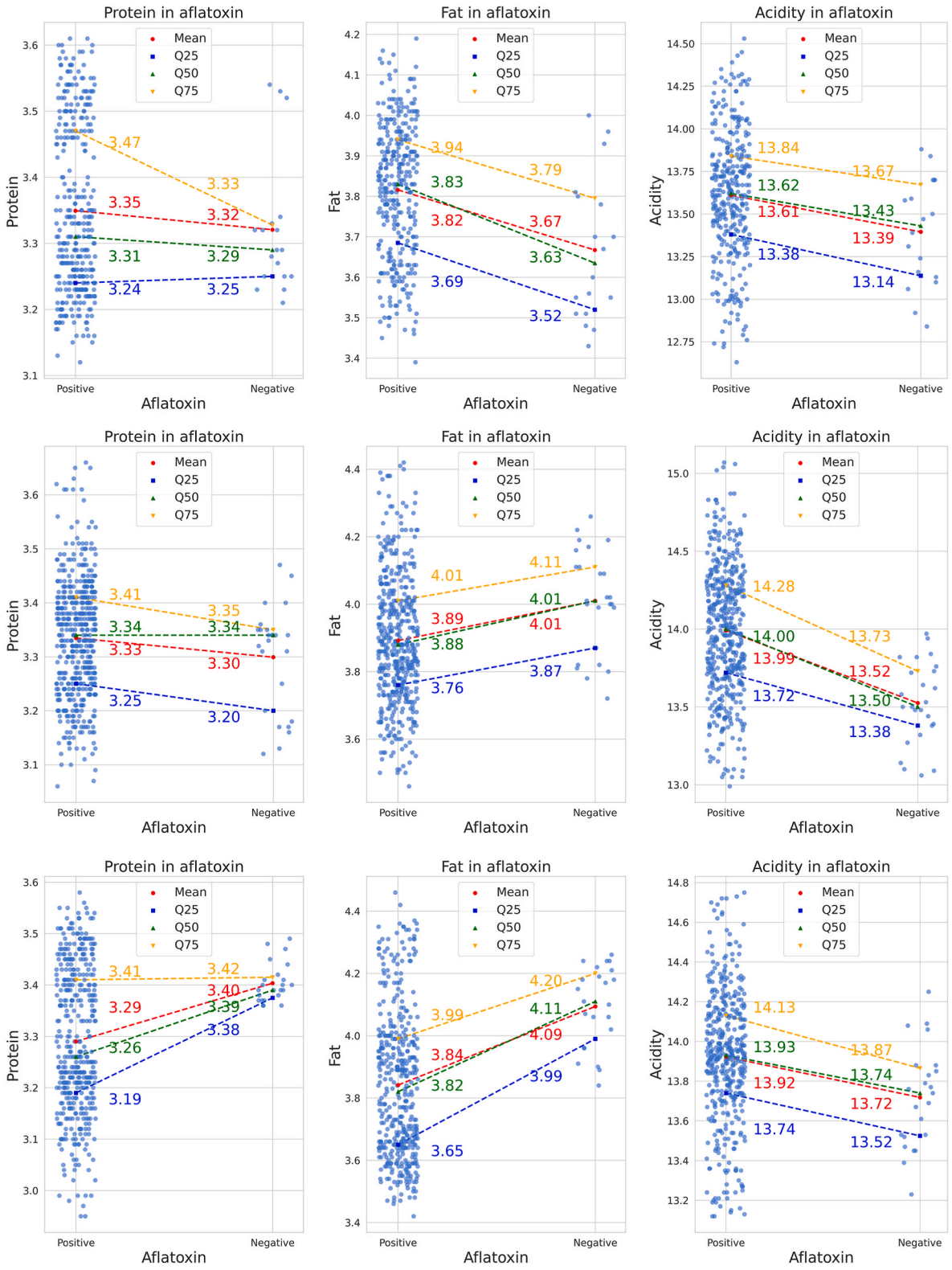
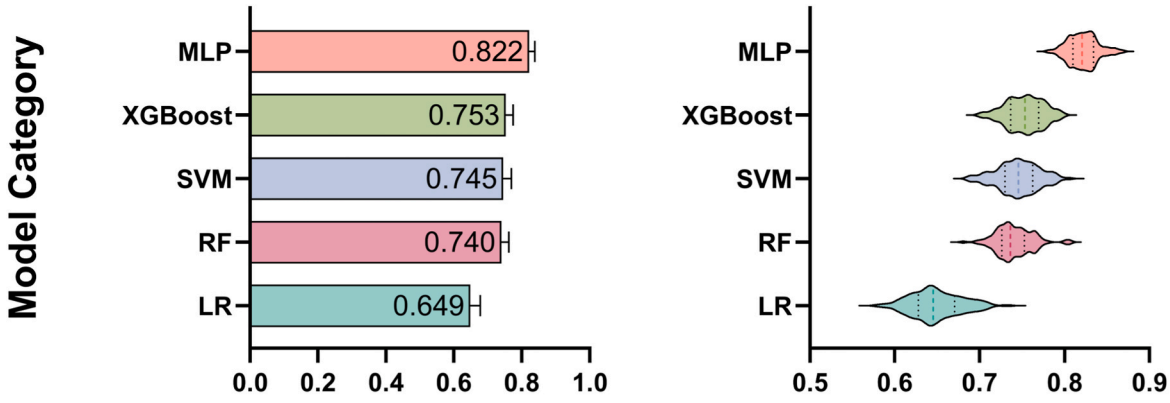


Fig. 3. Scatter plots of protein, fat, and acidity, stratified by AFM_1 status (positive vs. negative) across regions A, B, and C. Horizontal lines represent quartiles.

acidity, suggesting potential influences of microbial activity on the presence of toxin. These findings highlight that regional production environments strongly influence milk composition and AFM_1 distribution.

In Fig. 3a (Region A), 18 negative samples were detected from 407 raw milk samples; in Fig. 3b and c (Region B and Region C), 25 and 23 negative samples were detected from 598 to 523 raw milk samples, respectively. We compared protein, fat, and acidity in these samples and analyzed trends using mean, 25th, 50th, and 75th percentiles (Fig. 3). In Region A, negative AFM_1 have a small number of protein and fat values, whereas the fat in Region B and the protein and fat in Region C are distributed in higher ranges. There is no obvious difference in the



(a) Comparison of the overall accuracy of five machine learning models with standard deviation on the dataset.

(b) Violin plot of the overall accuracy of the model for 100 experiments.

Fig. 4. Comparison of model performance. The white dot represents the median, and the thick bar represents the interquartile range.

Table 3
The variance of the evaluation results of each model under 100 experiments. The 0 after "_" indicates a positive sample, and 1 indicates a negative sample. The model with the lowest variance in each indicator is underlined.

Model	Accuracy	Precision_0	Recall_0	F1-score_0	Precision_1	Recall_1	F1-score_1
LR	0.030	0.045	0.047	0.031	0.037	0.050	0.033
RF	0.022	0.037	0.041	0.025	0.036	0.042	0.024
SVM	0.024	0.040	0.039	0.023	0.033	0.039	0.027
XGBoost	0.021	0.033	0.038	0.025	0.036	0.041	0.021
MLP	0.017	0.046	0.053	0.032	0.042	0.051	0.034

protein distribution in Region B. It should also be noted that negative AFM_1 from all three regions exhibit low acidity. This suggests that aflatoxin content is significantly influenced by regional differences but may cause slightly lower acidity, likely due to microbial differences in raw milk.

3.2. Comparison of model prediction results

We conducted comparative experiments using LR, RF, SVM, XGBoost, and MLP to evaluate their performance in AFM_1 classification. The overall test-set accuracies achieved by the five models are presented in Fig. 4, while Table 3 summarizes the variance of each evaluation metric across 100 independent runs. Among all models, MLP achieved the highest accuracy (82.2%) with the lowest variance (0.017), demonstrating superior predictive performance and excellent stability. The optimized MLP architecture consisted of a 15-dimensional input layer, one hidden layer with 128 units, and a single-node output layer representing the probability of AFM_1 positivity. The model used a learning rate of 0.001, batch size of 16, and 20 training epochs. Given the limited number of features and samples, a relatively shallow architecture provided the best balance between model capacity and overfitting risk.

Across the remaining models, RF, SVM, and XGBoost exhibited similar performance, while LR consistently performed the worst—consistent with the PCA results (Fig. 3), which showed no strong linear separability in the data, making LR less suitable for this task. As shown in Fig. 5, all models predicted positive samples (AFM_1-positive = 0) more accurately than negative samples (AFM_1-negative = 1). This trend can be attributed to the data imbalance, as the larger number of positive samples enabled the models to more effectively learn their distributional characteristics.

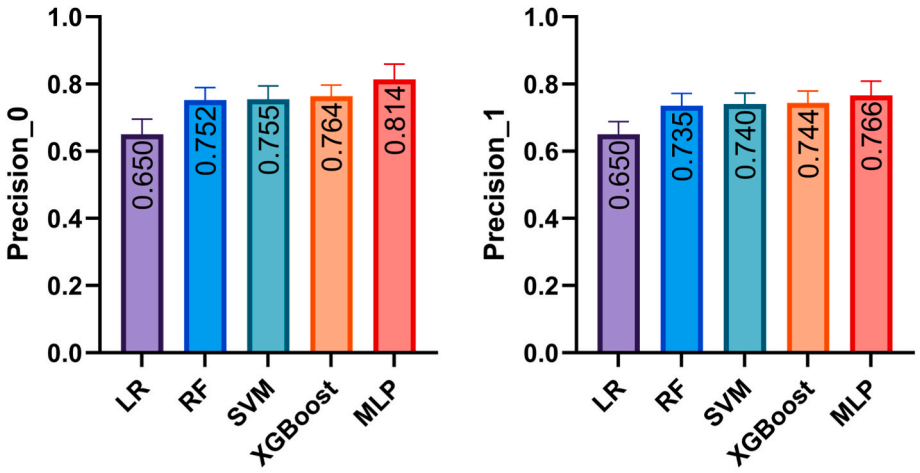
To further assess model stability, we conducted 100 repeated experiments and performed paired t-tests. MLP significantly outperformed LR ($p < 0.01$). Although the precision of MLP exceeded that of RF, SVM, and XGBoost, the differences were not statistically significant ($p > 0.05$), suggesting that tree-based and kernel-based models achieved comparable average precision. Importantly, recall of negative samples (class 1) is a critical metric for AFM_1 screening, as it reflects the model's ability to avoid false negatives—situations where unqualied milk could be misclassified as safe (Juba and Le, 2019). MLP achieved the highest recall for the negative class (83.2%), outperforming XGBoost by 6.3 percentage points (0.832 vs. 0.769). This indicates that fewer than two in ten AFM_1-negative samples were misclassified, making MLP the most reliable model for risk-averse industrial screening scenarios. Moreover, MLP achieved the highest F1-scores for both positive and negative classes, further demonstrating its balanced performance.

To further assess the generalizability and robustness of the model beyond the primary dataset, we evaluated the best-performing MLP model (achieving 87 % accuracy over 100 training runs) on an independent raw milk dataset collected in 2018. This dataset contains 8634 AFM_1 positive samples and 137 negative samples, representing a highly imbalanced real-world scenario.

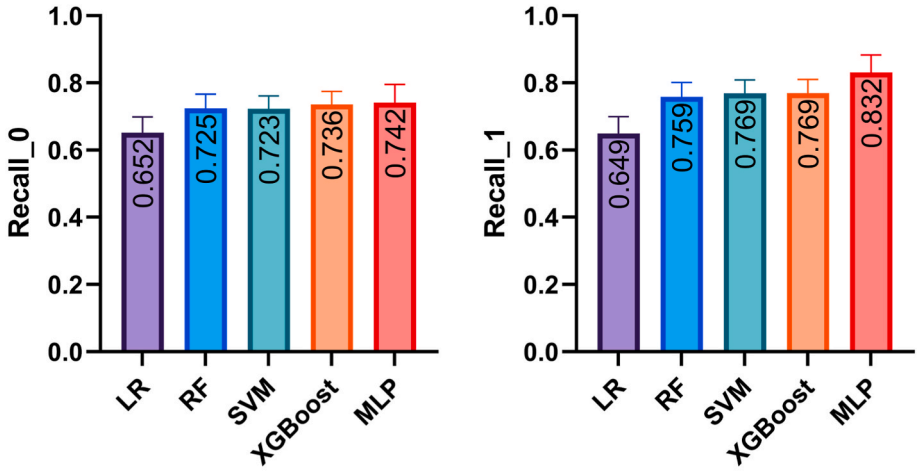
The confusion matrix of the external validation is presented in Fig. 6, and the corresponding performance metrics are summarized in Table 4. As shown in Fig. 6, the model correctly identified 104 negative samples and 7841 positive samples, while misclassifying 793 samples as false negative and 33 samples as false positive (see Fig. 6).

Table 4 indicates that the positive class exhibited a high recall (75.91%) but a relatively low precision (11.59%), whereas the negative class achieved both high precision (99.58%) and high F1-score (0.9500). This pattern highlights a well-known effect in imbalanced classification. The number of actual positive samples is extremely small (137), even a moderate number of false positives (FP = 793) leads to a substantial decline in precision. However, the model successfully captured the majority of true AFM_1 negative samples, resulting in strong recall performance, which is the most important metric for food safety early warning systems.

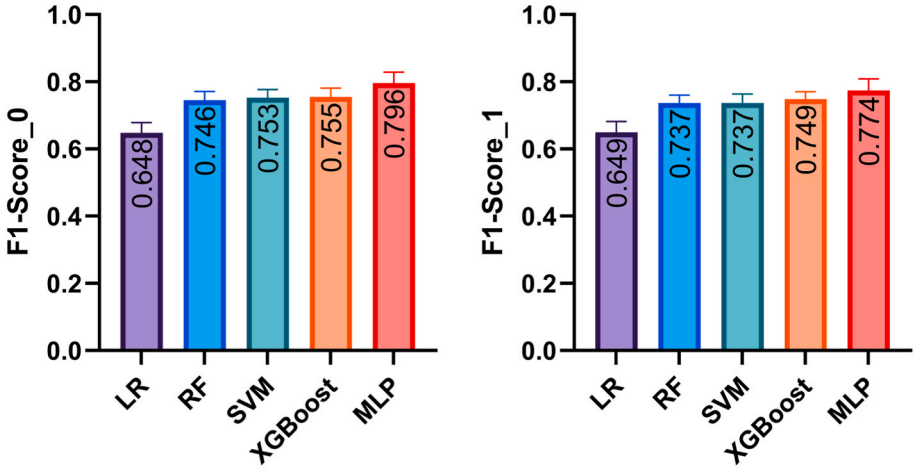
This behavior is consistent with the evaluation results obtained from



(a) Precision of the five models on the test set



(b) Recall of the five models on the test set



(c) F1-score of the five models on the test set

Fig. 5. Performance comparison of the models based on Precision, Recall, and F1-score, reported separately for each class label. Class 0 corresponds to qualified samples with AFM_1 (<0.05 μg/kg), and Class 1 corresponds to unqualified samples with AFM_1 (≥ 0.05 μg/kg). Error bars denote the standard deviation across 100 independent training runs.

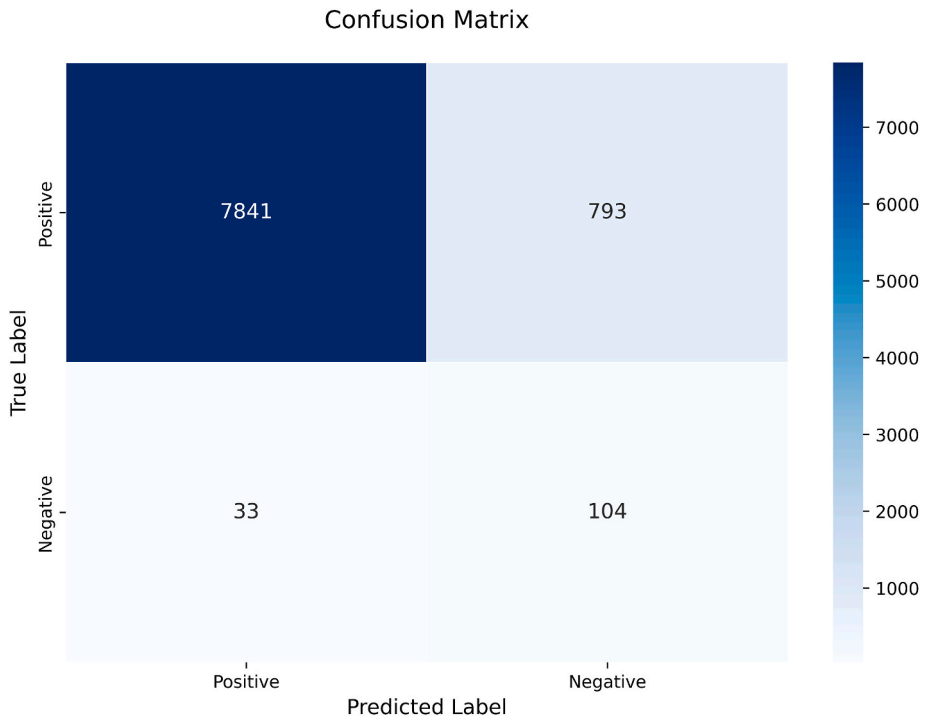


Fig. 6. Confusion matrix of the MLP model evaluated on the independent 2018 external validation dataset.

Table 4
Performance metrics of the MLP model on the independent 2018 external validation dataset.

Class	Precision	Recall	F1-score
1 (AFM_1 negative)	0.1159	0.7591	0.2012
0 (AFM_1 positive)	0.9958	0.9082	0.9500

the validation sets of other datasets. The high recall and low false-negative rate (33 missed positives out of 8771 samples, 0.38%) demonstrate the model's ability to preserve its decision boundary even when applied to data from a different year. From a practical perspective, ensuring that contaminated milk is not missed is far more critical than reducing false alarms; therefore, the model's tendency to favor sensitivity over precision is acceptable in industrial risk control scenarios.

3.3. Feature correlation analysis

To further interpret how the optimal MLP model behaves on unseen data, SHAP analysis was conducted on 1000 samples drawn from the independent 2018 dataset in Fig. 7. Unlike traditional feature importance metrics, SHAP values provide both the magnitude and the direction of each feature's influence on the model's output. Therefore, in addition to comparing mean absolute SHAP values, we also examined the full distribution of SHAP attributions including variance, percentile spread, and asymmetry to characterize how different features affect AFM_1 predictions across diverse samples.

Among all features, lactose and viscosity showed the largest SHAP variances (std = 0.40 and 0.55, respectively) and broad value ranges. The value of lactose is from −1.91 to +0.36. The value of viscosity is from −2.22 to +0.77. The shape of these distributions is highly asymmetric, with large negative tails, indicating that lower lactose levels and lower viscosity substantially increase AFM_1 positive predictions. These features not only display strong average importance but also high heterogeneity, meaning their influence varies significantly among regions and seasons. This behavior aligns with known biochemical signatures of AFM_1 contamination, where lactose degradation and viscosity alterations often accompany microbial activity or poor feed quality.

Psychrophiles (std = 0.47) and colony count (std = 0.34) have wide positive SHAP ranges, indicating that elevated microbial loads strongly push predictions toward AFM_1 positivity. The SHAP distributions of both features exhibit fat positive tails, meaning a subset of samples is highly sensitive to microbial contamination. This aligns with literature showing that poor feed storage and higher microbial growth correlate with increased aflatoxin exposure.

Milk temperature, acidity, fat, and protein exhibit moderate SHAP magnitudes but relatively symmetrical distributions around zero. These features consistently contribute to model decisions and their effect direction varies depending on the sample. Such as the value of acidity is from −1.83 to +0.56 and the value of fat is from −0.84 to +0.50. This pattern reflects physiochemical variability across regions and herds, consistent with the findings in Fig. 3 where regional fat or protein levels differed but did not always show monotonic associations with AFM_1.

A key observation is that Province and County ranked among the top six most influential features, with considerable SHAP dispersion (Province: mean = 0.36, std = 0.23, range = −0.74 to +0.77 and County: mean = 0.28, std = 0.41, range = −0.67 to +1.39). Unlike physicochemical features whose SHAP values often show symmetric distributions, geographic features exhibit distinctly asymmetric and bimodal patterns, indicating that they consistently increase predicted AFM_1 risk (positive SHAP clusters). And others consistently decrease it (negative SHAP clusters). This reflects true spatial heterogeneity in AFM_1 contamination, driven by differences in climate, humidity, crop harvesting and feed storage practices.

Overall, the SHAP analysis demonstrates that the MLP model forms its predictions through a stable and biologically plausible integration of physicochemical composition, microbial load, and region-specific factors, confirming that its decision-making process generalizes across datasets and reflects genuine patterns of AFM_1 risk rather than dataset-specific artifacts.

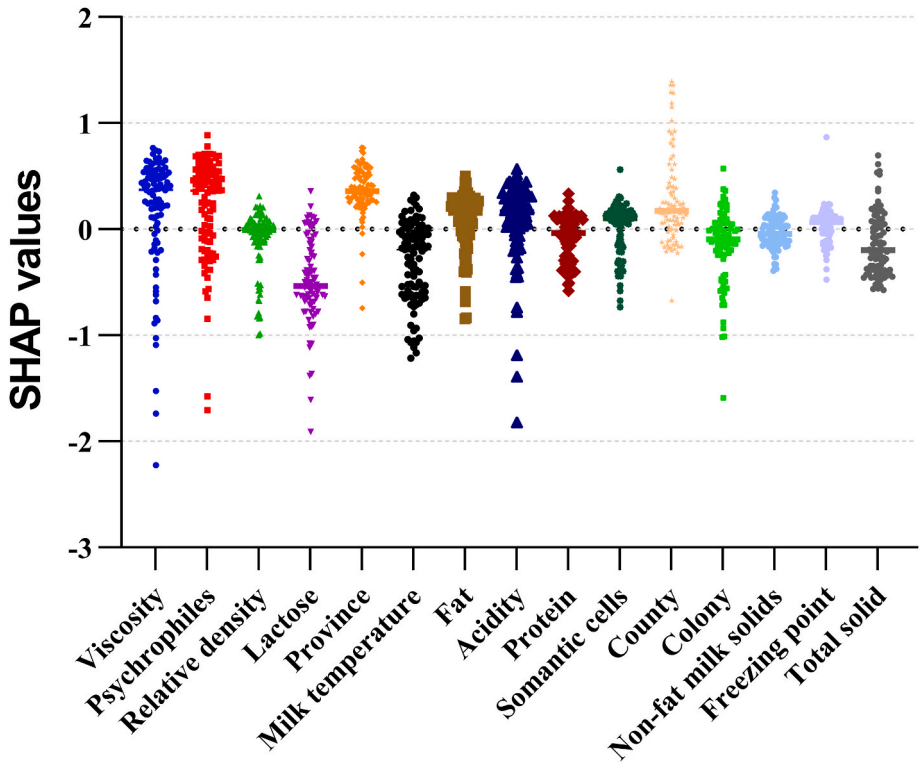


Fig. 7. SHAP summary plot showing individual feature contributions across samples. Each point represents one sample's SHAP value; features are ordered by average contribution to model out.

3.4. Limitations and future research

3.4.1. Limitations of the study

Despite the promising results demonstrated by the MLP model in predicting AFM_1 contamination, several limitations inherent to the study design and data characteristics must be acknowledged.

First, the limited number of naturally unqualified samples remains a primary constraint. Although the dataset contains over 40,000 records, only approximately 500 samples were identified as non-compliant (unqualified). To address the severe class imbalance, this study employed a random under-sampling strategy on the majority class. Although this created a balanced training set, it inevitably resulted in the exclusion of a large volume of compliant samples, potentially discarding valuable information regarding the variance of normal milk physiology. Consequently, the model may not fully capture the complete spectrum of safe raw milk profiles, leading to a potential risk of false positives when deployed on data distributions that differ from the sampled subset. In terms of input features, the prediction is based on indirect associations. The model infers toxin risk based on changes in basic physicochemical indicators (e.g., protein, viscosity) rather than measuring the toxin itself. Although SHAP analysis confirmed strong correlations, these relationships are correlational rather than causal. Therefore, in scenarios where AFM_1 contamination occurs without significantly altering the bulk physicochemical properties of the milk (e.g., low-level contamination), the model's sensitivity might be reduced.

3.4.2. Rationale for model selection

This study adopted MLP as the primary modeling method rather than deeper or more complex architectures such as Transformers (Vaswani et al., 2017) or ResNets (He et al., 2016). This decision was based on two practical considerations aligned with the data limitations described above. The input space consists of only 15 numerical features. Deep architectures like Transformers are designed to model complex dependencies in high-dimensional data (e.g., text or images). For low-dimensional tabular data, the structural advantages of attention mechanisms are limited, and simpler models often yield comparable or superior performance with lower computational cost. Besides, the scarcity of positive (unqualified) samples is insufficient to support stable training of high-capacity models. Deep models typically require large amounts of labeled data to converge without overfitting. Given the limited number of non-compliant samples, applying large-scale deep learning models would lack statistical justification and increase the risk of learning spurious correlations.

3.4.3. Implications for industrial application

To translate this predictive model into industrial practice, it is crucial to define its operational role correctly. As highlighted by food safety protocols, a machine learning model can help guide monitoring efforts but can never completely replace traditional analytical detection methods (such as ELISA or HPLC).

The proposed method is best implemented as a primary screening tool or an early warning system within a multi-tiered quality control framework. By continuously processing low-cost routine measurements (protein, fat, etc.) that are already being collected, the model can assign a risk score to each batch of raw milk. Batches flagged as high-risk can then be prioritized for immediate and rigorous chemical testing. This tiered approach optimizes the allocation of expensive testing resources and improves the overall efficiency of the safety monitoring chain. Companies must calibrate the decision threshold based on their specific economic and safety risk tolerance. A strict high-recall threshold is recommended to minimize the risk of missing unqualified samples, even if it incurs a higher rate of confirmatory tests. After that, a mechanism for dynamic model updates is essential. As new verified lab results become available, they should be fed back into the system to retrain the model, allowing it to adapt to seasonal changes in raw milk characteristics.

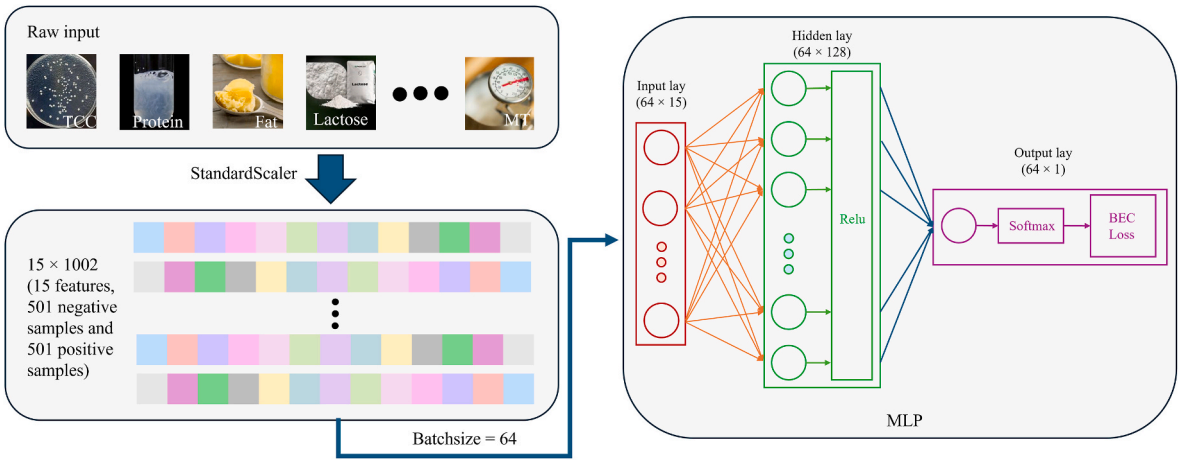


Fig. 8. Overview of the research methodology (TCC means total colony count. T means milk temperature).

3.5. Future research directions

Future work should focus on overcoming the data and algorithmic limitations identified above. Firstly, the most critical next step is to expand the collection of real positive data through cross-factory and multi-laboratory collaborations. Establishing a representative, multi-regional AFM_1 contamination dataset will allow for robust external validation and improve the model's ability to generalize across different supply chains. With the accumulation of more data, future research should explore data-efficient algorithms specifically designed for tabular learning, such as TabNet (Arik and Pfister, 2021), FT-Transformer (Gorishniy et al., 2021), or XGBoost variants. Additionally, semi-supervised or transfer learning approaches could be investigated to take advantage of the vast amounts of unlabeled or compliant data available in dairy factories, potentially providing robust feature representations that improve toxin detection performance even with limited positive labels.

4. Conclusion

This study demonstrates that routinely collected physicochemical indicators can be effectively leveraged for AFM_1 risk assessment using machine learning. Among the five models evaluated, the MLP achieved the most stable performance, with accuracy above 80% and strong recall for AFM_1-negative samples. External validation using an independent 2018 dataset confirmed the model's robustness and low false-negative rate, supporting its suitability for practical safety monitoring. SHAP interpretation further showed that microbial factors, chemical traits, and geographical origins collectively shape AFM_1 predictions, indicating that the model captures biologically meaningful patterns rather than dataset-specific noise. Although the model cannot replace reference analytical methods, it can serve as an efficient pre-screening or early-warning tool to prioritize batches for confirmatory testing. Broader multi-regional validation, expanded unqualified datasets and regular model updates will be essential before large-scale industrial deployment.

In conclusion, as illustrated in Fig. 8 this work provides an interpretable and practical framework for enhancing AFM_1 surveillance in dairy production and offers a promising direction for integrating AI-based decision support into food safety management.

Credit author statement

Haohan Ding: Writing – review & editing, Writing – original draft, Conceptualization. Long Wang: Writing – review & editing, Writing – original draft, Validation.
Xiaodong Song: Writing – review & editing, Funding acquisition.
Xiaohui Cui: Writing – review & editing, Supervision.
David I. Wilson: Writing – review & editing, Software.
Wei Yu: Writing – review & editing, Validation.
Cheng Zhang: Writing – review & editing, Resources.
Guanjun Dong: Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by the National Key Research and Development Program of China (2024YFE0199500). The authors would like to express their sincere gratitude to Professor Brent R. Young from the University of Auckland for his invaluable advice, guidance, and support

Data availability

The data that has been used for this study is commercially sensitive, and is confidential.

References

Šašić, S., Ozaki, Y., 2001. Short-wave near-infrared spectroscopy of biological fluids. 1. Quantitative analysis of fat, protein, and lactose in raw milk by partial least-squares regression and band assignment. Anal. Chem. 73 (1), 64–71.
Arik, S.Ö., Pfister, T., 2021. Tabnet: attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6679–6687.
Bao, K., Liu, X., Xu, Q., Su, B., Liu, Z., Cao, H., Chen, Q., 2021. Nanobody multimerization strategy to enhance the sensitivity of competitive ELISA for detection of ochratoxin A in coffee samples. Food Control 127, 108167.
Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for hyper-parameter optimization. Adv. Neural Inf. Process. Syst. 24.
Bilandžić, N., Varga, I., Varenina, I., Solomun Kolanović, B., Božić Luburić, Đ., Đokić, M., Sedak, M., Cvetnić, L., Cvetnić, Ž., 2022. Seasonal occurrence of Aflatoxin M1 in raw milk during a five-year period in Croatia: dietary exposure and risk assessment. Foods 11 (13), 1959.
Bridle, J., 1989. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. Adv. Neural Inf. Process. Syst. 2.
Castano-Duque, L., Vaughan, M., Lindsay, J., Barnett, K., Rajasekaran, K., 2022. Gradient boosting and bayesian network machine learning models predict aflatoxin and fumonisin contamination of maize in Illinois–First USA case study. Front. Microbiol. 13, 1039947.
Chen, T., 2016. XGBoost: a Scalable Tree Boosting System. Cornell University.
Clark, J.L., Barbano, D.M., Dunham, C.E., 1989. Comparison of two methods for determination of total solids content of milk: collaborative study. J. Assoc. Off. Anal. Chem. 72 (5), 712–718.

de Freitas, A.G.M., Minho, L.A.C., de Magalhães, B.E.A., Dos Santos, W.N.L., Santos, L.S., de Albuquerque Fernandes, S.A., 2021. Infrared spectroscopy combined with random forest to determine tylosin residues in powdered milk. Food Chem. 365, 130477.
Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A., 2021. Revisiting deep learning models for tabular data. Adv. Neural Inf. Process. Syst. 34, 18932–18943.
Greenacre, M., Groenen, P.J., Hastie, T., d'Enza, A.I., Markos, A., Tuzhilina, E., 2022. Principal component analysis. Nat. Rev. Methods Primers 2 (1), 100.
Harding, F., 1995. Adulteration of Milk. Milk Quality. Springer, pp. 60–74.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
Hernández-Martínez, R., Navarro-Blasco, I., 2015. Surveillance of aflatoxin content in dairy cow feedstuff from Navarra (Spain). Anim. Feed Sci. Technol. 200, 35–46.
Hossin, M., Sulaiman, M.N., 2015. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process 5 (2), 1.
Jiang, Y., Chen, S., Bian, B., Li, Y., Sun, Y., Wang, X., 2021. Discrimination of tomato maturity using hyperspectral imaging combined with graph-based semi-supervised method considering class probability information. Food Anal. Methods 14 (5), 968–983.
Juba, B., Le, H.S., 2019. Precision-Recall versus Accuracy and the Role of Large Data Sets. Proceedings of the AAAI Conference on Artificial Intelligence 33 (01), 4039–4048. https://doi.org/10.1609/aaai.v33i01.33.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
Liashchynskyi, P., Liashchynskyi, P., 2019. Grid Search, Random Search, Genetic Algorithm: a Big Comparison for NAS. *arXiv preprint arXiv:1912.06059*.
Lu, M., Shiau, Y., Wong, J., Lin, R., Kravis, H., Blackmon, T., Pakzad, T., Jen, T., Cheng, A., Chang, J., 2013. Milk spoilage: methods and practices of detecting milk quality. Food Nutr. Sci. 4 (7), 113–123.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.
Mansuri, S.M., Chakraborty, S.K., Mahanti, N.K., Pandiselvam, R., 2022. Effect of germ orientation during Vis-NIR hyperspectral imaging for the detection of fungal contamination in maize kernel using PLS-DA, ANN and 1D-CNN modelling. Food Control 139, 109077.
Mao, A., Mohri, M., Zhong, Y., 2023. Cross-entropy loss functions: theoretical analysis and applications. In: International Conference on Machine Learning. pmlr, pp. 23803–23828.
McKinnon, K.M., 2018. Flow cytometry: an overview. Curr. Protoc. Im. 120 (1), 5.1. 1-5.1. 11.
Mudannayake, A., Karunarathne, S., Jayasooriya, P.W., Nanayakkara, D., Abesooriya, A., Silva, S., Fernando, R., 2024. Occurrence of aflatoxin M1 in cheese products commonly available in Sri Lankan market. Heliyon 10 (15), e35155.
Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814.
Oliveira, G.B.d., Favarin, L., Luchese, R.H., McIntosh, D., 2015. Psychrotrophic bacteria in milk: how much do we really know? Braz. J. Microbiol. 46 (2), 313–321.
Omar, S.S., 2016. Aflatoxin M1 levels in raw milk, pasteurised milk and infant formula. Italian journal of food safety 5 (3), 5788.
Peña-Rodas, O., Martinez-Lopez, R., Hernandez-Rauda, R., 2018. Occurrence of Aflatoxin M1 in cow milk in El Salvador: results from a two-year survey. Toxicol. Rep. 5, 671–678.
Picciano, M.F., 2001. Nutrient composition of human milk. Pediatr. Clin. 48 (1), 53–67.
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. 31.
Ribeiro, M.T., Singh, S., Guestrin, C., 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.
Rodrigues, I., Naehrer, K., 2012. A three-year survey on the worldwide occurrence of mycotoxins in feedstuffs and feed. Toxins 4 (9), 663–675.
Singh, H., Singh, S., Bhardwaj, S.K., Kaur, G., Khatri, M., Deep, A., Bhardwaj, N., 2022. Development of carbon quantum dot-based lateral flow immunoassay for sensitive detection of aflatoxin M1 in milk. Food Chem. 393, 133374.
Smoke, T., Smoking, I., 2004. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, vol. 1. IARC, Lyon, pp. 1–1452.
Soper, D.S., 2021. Greed is good: rapid hyperparameter optimization and model selection using greedy k-fold cross validation. Electronics 10 (16), 1973.
Tadesse, S., Berhanu, T., Woldegiorgis, A.Z., 2020. Aflatoxin M1 in milk and milk products marketed by local and industrial producers in Bishoftu town of Ethiopia. Food Control 118, 107386.
Tsenkova, R., Atanassova, S., Kawano, S., Toyoda, K., 2001. Somatic cell count determination in cow's milk by near-infrared spectroscopy: a new diagnostic tool. J. Anim. Sci. 79 (10), 2550–2557.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Wang, J., Zhu, L., Zhang, W., Wei, Z., 2019. Application of the voltammetric electronic tongue based on nanocomposite modified electrodes for identifying rice wines of different geographical origins. Anal. Chim. Acta 1050, 60–70.
Wu, X., Zhu, J., Wu, B., Zhao, C., Sun, J., Dai, C., 2019. Discrimination of Chinese liquors based on electronic nose and fuzzy discriminant principal component analysis. Foods 8 (1), 38.
Xiong, J., Peng, L., Zhou, H., Lin, B., Yan, P., Wu, W., Liu, Y., Wu, L., Qiu, Y., 2020. Prevalence of aflatoxin M1 in raw milk and three types of liquid milk products in central-south China. Food Control 108, 106840.
Yen, S.-J., Lee, Y.-S., 2009. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst. Appl. 36 (3), 5718–5727.
Zheng, N., Sun, P., Wang, J., Zhen, Y., Han, R., Xu, X., 2013. Occurrence of aflatoxin M1 in UHT milk and pasteurized milk in China market. Food Control 29 (1), 198–201.