Journal of Food Engineering 426 (2027) 113386

Contents lists available at ScienceDirect

Journal of Food Engineering

journal homepage: www.elsevier.com/locate/jfoodeng

Machine learning-based analysis of process factors associated with quality variability in industrial probiotic manufacturing

Yunjeong Jang [a], BoYoung Choi [b], Roneel V. Sharan [c], Haeseong Park [b], Yohwan Kim [b], Jin Seok Moon [b,*], Minji Lee [a,*]

[a] Department of Biomedical Software Engineering, The Catholic University of Korea, Bucheon-si, Gyeonggi, 14662, South Korea
[b] Ildong Bioscience, Pyeongtaek-si, Gyeonggi, 17957, South Korea
[c] School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, United Kingdom

ARTICLE INFO

Keywords:
Machine learning
Probiotic manufacturing
Quality variability
Cell viability
Feature importance

ABSTRACT

Ensuring both high productivity and viable cell survival remains a major challenge in industrial-scale probiotic manufacturing due to the multistage and highly variable nature of fermentation and downstream processing. This study presents an exploratory, machine learning-based framework for analyzing batch-to-batch quality variability in the production of an L. rhamnosus-based probiotic, using data from 178 independent production batches together with process variables measured across the entire manufacturing pipeline, from seed culture through centrifugation, freeze-drying, and sieving. Three quality attributes — total production, viable cell count (VCC) reduction relative to theoretical values, and VCC change after one week of storage — were each dichotomized into below- and above-mean groups, framing batch-to-batch quality variation as a binary classification problem. XGBoost models were implemented to predict each quality attribute; the predictions of total production and VCC reduction achieved higher classification performance than the prediction of VCC change after one week. Feature importance analysis identified cryoprotectant volume, centrifugation-related variables (e.g., number of discharges), and pressure-related variables as prominent predictors, although their stability varied across quality attributes and sensitivity analyses. To verify the robustness of the results, a range of sensitivity analyses was additionally conducted, including comparisons across missing-data handling thresholds and classification cutoff definitions (mean vs. median). These analyses showed that the prediction of total production remained comparatively stable, whereas the predictions of VCC reduction and VCC change were somewhat sensitive to these methodological choices. These findings offer exploratory, process-level insights into predictive quality assessment based on routinely collected process data, while also indicating that their interpretation warrants considerable caution.

1. Introduction

Probiotics have attracted increasing scientific and commercial interest due to their roles in modulating gut microbiota, enhancing immune function, and supporting metabolic health (Quigley, 2011; Abdul Manan, 2025; Reddyvari et al., 2025; Esposito et al., 2025). Among various probiotic strains, *Lactobacillus rhamnosus* is widely used owing to its tolerance to gastric acid and bile salts, as well as its ability to adhere to the intestinal mucosa, contributing to its well-documented health benefits (Reddyvari et al., 2025; Pagnini et al., 2018; Ghiasi et al., 2023; Yang et al., 2025). As a result, industrial demand for large-scale and consistent production has increased significantly. However, maintaining consistent product quality remains challenging due to the complex and multistage nature of probiotic manufacturing processes, which involve tightly coupled fermentation and downstream operations (Onel et al., 2018). Variability in process conditions across stages can lead to substantial differences in production yield and cell viability, highlighting the need for systematic approaches to analyze and predict quality variability in industrial production systems (Onel et al., 2018).

In probiotic manufacturing, product quality is primarily defined by the viable cell count (VCC), expressed in colony-forming units (CFU), which serves as a key indicator of functional efficacy (Pagnini et al., 2018; Ghiasi et al., 2023). Industrial quality assessment typically involves measuring VCC immediately after downstream processing, as well as after a defined storage period to evaluate survival stability (Carvalho et al., 2004; Morgan et al., 2006). Despite standardized

* Corresponding authors.
E-mail addresses: jyj20107@catholic.ac.kr (Y. Jang), benmartim@ildong.com (B. Choi), roneel.sharan@essex.ac.uk (R.V. Sharan), haeseong@ildong.com (H. Park), yohk504@ildong.com (Y. Kim), moonjs@ildong.com (J.S. Moon), minjilee@catholic.ac.kr (M. Lee).

https://doi.org/10.1016/j.jfoodeng.2026.113386
Received 20 June 2026; Received in revised form 13 October 2026; Accepted 17 October 2026
Available online 29 October 2026
0260-8774/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

evaluation protocols, substantial batch-to-batch variability in VCC is frequently observed in practice (Doran, 2013; Rathore et al., 2021), arising from complex interactions among process conditions across multiple stages that are difficult to disentangle using conventional approaches. These challenges highlight the need for data-driven modeling approaches that can capture multivariate relationships and enable predictive analysis of quality variability in industrial production systems (Onel et al., 2018).

Probiotic manufacturing involves a sequence of unit operations, including fermentation, centrifugation, freeze-drying, sieving, and storage, in which process conditions interact with the physiological state of cells to influence final product quality. During these stages, cells are exposed to various physical and chemical stresses that can induce cellular damage and reduce viability (Carvalho et al., 2004; Morgan et al., 2006; Papadimitriou et al., 2016), with variability in production yield and viable cell count arising from complex interactions among process parameters such as medium composition, temperature, pH, agitation, and fermentation duration (Doran, 2013; Rathore et al., 2021). Because these effects accumulate and interact across stages, analyses based on individual parameters remain insufficient to capture the full complexity of industrial production systems (Doran, 2013; Rathore et al., 2021); conventional approaches often focus instead on specific stages or predefined variables (Miecznikowski et al., 2026; Breig and Luti, 2021), overlooking cross-stage dependencies and emergent effects arising from the coupling of upstream and downstream processes. A process-wide analytical framework integrating information across the entire manufacturing pipeline is therefore needed to effectively characterize and manage quality variation in probiotic production.

Data-driven modeling approaches offer a promising solution in this context, enabling the integration of multistage process data and capturing complex, nonlinear interactions in industrial bioprocess systems. Machine learning (ML) approaches, in particular, have gained attention for analyzing patterns and predicting quality in biological manufacturing systems as large-scale process data have become increasingly available (Onel et al., 2018; Mondal et al., 2023; Pandey et al., 2026; Khanashyam et al., 2025; Adli et al., 2025), with tree-based ensemble models being well suited for bioprocess data due to their ability to capture nonlinear relationships and interactions among variables (Chen and Guestrin, 2016; Paul et al., 2025; Celades-Martínez et al., 2026; Ageh et al., 2024; Saha et al., 2025) and to identify key process variables beyond prediction (Lundberg and Lee, 2017; Kırbaş et al., 2019). However, few ML studies have integrated data across multiple processing stages to assess final-product quality (Al Hagbani et al., 2025; Tulsyan et al., 2018; Melcher et al., 2015).

A further limitation concerns the scale and origin of the data underlying most existing studies: much of the current evidence on process factors affecting probiotic viability during freeze-drying and storage is derived from laboratory- or pilot-scale experiments conducted under controlled conditions, focusing on isolated factors (Miecznikowski et al., 2026; Al Hagbani et al., 2025; Ge et al., 2024; Buahom et al., 2023), and whether such relationships extend to the scale and operational variability of industrial production — spanning upstream culture, downstream processing, and post-processing storage within a single manufacturing workflow — remains largely unexamined.

Here, we present an exploratory, data-driven analysis of industrial-scale probiotic manufacturing to investigate how process conditions are associated with relative inter-batch quality variability across the production pipeline. The motivation for this study was the lack of process-wide evidence showing how routinely recorded variables across multiple manufacturing stages are associated with quality variability under industrial operating conditions. Using real-world process data from the manufacturing of an *L. rhamnosus*-based product, we examined three quality attributes: total production yield, reduction in VCC relative to theoretical values, and VCC change during one-week post-processing storage. Accordingly, the major objectives were to evaluate the predictability of these three quality attributes, identify the process variables associated with their inter-batch variability, and assess the robustness of the identified patterns. XGBoost-based classification models were employed to capture nonlinear relationships among these variables, and the resulting feature importance was used to examine which process variables contributed most strongly to the predictions. To assess the robustness of these findings, we further conducted a series of sensitivity and comparative analyses. This study offers exploratory, process-level insights into quality variability in industrial probiotic manufacturing, illustrating how routinely collected process data may inform predictive quality assessment and process understanding in large-scale bioprocessing.

The principal novelties of this study are summarized as follows:

- Process-wide integration of routinely collected variables spanning seed culture, main culture, concentration, cryoprotectant addition, and freeze-drying using data from 178 industrial-scale production batches.
- Joint evaluation of three complementary quality attributes representing production yield, processing-related viability reduction, and short-term post-processing viability change, with input variables restricted to those available at the intended prediction time for each task.
- Robustness assessment of predictive performance and feature-importance patterns using alternative classification cutoffs, preprocessing conditions, and baseline machine learning models.

2. Materials and methods

2.1. Data source and manufacturing process

Production data were obtained from an industrial probiotic manufacturing process operated by ILDONG Bioscience Co., Ltd. (Pyeongtaek-si, Gyeonggi-do, Republic of Korea), with each data sample recorded at the batch level.

L. rhamnosus IDCC 3201 was produced through a batch fermentation process followed by harvesting, concentration, freeze-drying, sieving, and storage. The overall manufacturing workflow is shown in Fig. 1, which consisted of sequential stages, including pre-culture, first seed culture, second seed culture, and main culture, followed by downstream processing. The main fermentation was conducted in a 50 m^3 industrial-scale bioreactor with a working volume of 20 kL under controlled conditions (37 °C, pH 5.7). Process variables, including pH, temperature, and agitation speed, were continuously monitored using a PLC-based control system and recorded at 1–5 min intervals. After fermentation, cells were harvested using a continuous centrifugation system and subjected to concentration and freeze-drying with the addition of cryoprotective agents.

A total of 196 distinct production batches were collected, with each batch representing one independent batch-level observation. No production batch was repeated. Total cell count (TCC) and VCC were measured in technical triplicate within each batch, and the mean of the three measurements was used as a single batch-level value. These technical replicates were not treated as independent samples.

2.2. Process data and variable construction

To ensure consistency and comparability across heterogeneous process measurements, raw manufacturing data were standardized prior to variable construction. Because the raw manufacturing records differed in sampling frequency, scale, and format across process steps, they were transformed into structured batch-level variables using predefined rules (Onel et al., 2018; Mählkvist et al., 2023).

Specifically, raw observations were summarized into representative variables such as end-point values for each stage (e.g., final pH) and

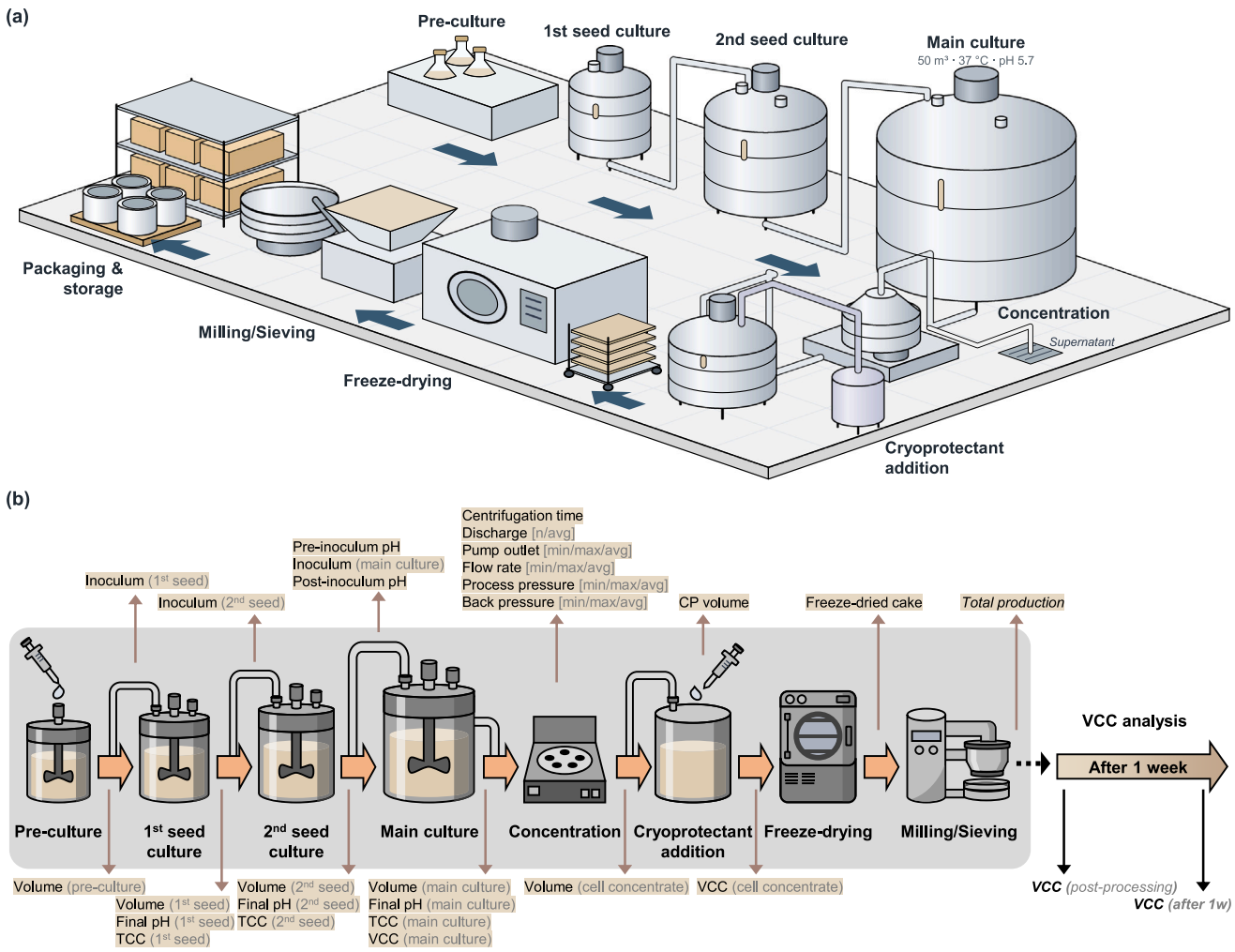


Fig. 1. Schematic overview of the industrial probiotic manufacturing process and the variables measured at each stage. (a) Plant-scale process flow, from pre-culture through the first and second seed cultures and the main culture, followed by concentration by centrifugation with removal of the supernatant, cryoprotectant addition, freeze-drying, milling/sieving, and packaging and storage. Pipes indicate the transfer of culture broth and cell concentrate between units, and arrows indicate the direction of material flow through the process. Equipment is shown schematically and is not drawn to scale. (b) Process stages and the corresponding measured variables included in this study. Variables highlighted with a background color were used as inputs to the classification models depending on the prediction target, whereas variables displayed in italics were either used to calculate the quality variables (post-processing VCC and VCC after 1 week) or served directly as a quality variable (total production). Total production was used as the target variable in the total production model and as an input variable only for VCC change prediction. Variables reported as [n/avg] or [min/max/avg] (e.g., discharge, pump outlet, flow rate, process pressure, and back pressure) represent distinct summary statistics derived from the same raw time-series measurements (see Section 2.2) and were counted as separate input variables.

count-based variables (e.g., discharge frequency). For the four measurements — pump outlet pressure, flow rate, process pressure, and back pressure — which were recorded at multiple time points within each process run, variable construction was performed to transform temporally resolved measurements into a fixed-length representation: descriptive statistics, specifically the minimum (min), maximum (max), and arithmetic mean (avg), were computed for each measurement (Onel et al., 2018). Time-series models such as autoregressive moving average (ARMA) models are widely used to characterize and forecast temporal dynamics in agricultural and industrial processes when long, regularly sampled records are available (Box et al., 2015; Afsharnia et al., 2018); for example, Afsharnia et al. (2018) fitted ARMA models to hourly sugarcane transport records comprising approximately 56,000 observations. In contrast, the four variables in this study were recorded at only one to four time points per batch (approximately 2.2–2.4 on average; Table 1), which is insufficient to estimate autoregressive or moving-average terms within a batch. Moreover, because the analytical unit of this study was the batch rather than the individual time point, each within-batch series had to be represented by a fixed number of batch-level features regardless of its length. We therefore adopted a feature-based representation in which each series is summarized by descriptive statistics (Onel et al., 2018; Christ et al., 2018). This transformation enabled heterogeneous time-series data to be integrated into a unified analytical dataset while preserving the operational relevance and interpretability of each variable. Fig. 1 shows the variables constructed from the measurements at each process stage.

2.3. Data preprocessing

2.3.1. Missing data filtering

Because the measured data contained missing values, data imputation was performed. Before imputation, variable- and sample-level exclusion based on missing rate was conducted in multiple steps to refine the dataset while preserving informative variables and samples (Oexner et al., 2025): (i) exclusion of variables with >50% missing rate; (ii) recalculation of missing rate and exclusion of samples with >25% missing rate; and (iii) recalculation of missing rate and exclusion of variables with >25% missing rate. The 50% threshold in step (i) was

Table 1
Summary of process variables and their use in predictive modeling.

Feature variable	Abbreviation	Unit	# recordings[a], raw/included	# observations[b], mean (SEM)	Description
1st seed culture inoculum	Inoculum (1st seed)	L	173/168	–	Volume of inoculum used to initiate the first seed culture
2nd seed culture inoculum	Inoculum (2nd seed)	L	178/173	–	Volume of inoculum used to initiate the second seed culture
Main culture inoculum	Inoculum (main culture)	L	172/167	–	Volume of inoculum used to initiate the main culture
Pre-culture volume	Volume (pre-culture)	mL	177/170	–	Culture broth volume measured at the end of the pre-culture stage
1st seed culture volume	Volume (1st seed)	L	178/173	–	Culture broth volume measured at the end of the first seed culture
2nd seed culture volume	Volume (2nd seed)	L	178/173	–	Culture broth volume measured at the end of the second seed culture
Main culture volume	Volume (main culture)	L	196/178	–	Culture broth volume measured at the end of the main culture
Cell concentrate volume[e]	Volume (cell concentrate)	kg	196/178	–	Mass of the cell concentrate
Pre-inoculation main culture pH	Pre-inoculation pH	pH	150/142	–	pH of the main culture medium immediately before inoculation with the seed culture
Post-inoculation main culture pH	Post-inoculation pH	pH	173/164	–	pH of the main culture medium immediately after inoculation with the seed culture
1st seed culture final pH	Final pH (1st seed)	pH	153/144	–	Final pH of the culture medium measured at the end of the first seed culture
2nd seed final pH	Final pH (2nd seed)	pH	149/145	–	Final pH of the culture medium measured at the end of the second seed culture
Main culture final pH	Final pH (main culture)	pH	145/138	–	Final pH of the culture medium measured at the end of the main culture
1st seed culture total cell count	TCC (1st seed)	10^8 cells/mL	178/172	3 (0)[c]	Total cell count per unit volume measured at the end of the first seed culture
2nd seed culture total cell count	TCC (2nd seed)	10^8 cells/mL	176/169	3 (0)[c]	Total cell count per unit volume measured at the end of the second seed culture
Main culture total cell count	TCC (main culture)	10^8 cells/mL	184/168	3 (0)[c]	Total cell count per unit volume measured at the end of the main culture
Main culture viable cell count[f]	VCC (main culture)	10^8 CFU/mL	194/176	3 (0)[c]	Viable cell count per unit volume at the end of the main culture
Cell concentrate viable cell count[f]	VCC (cell concentrate)	10^8 CFU/mL	196/178	3 (0)[c]	Viable cell count per unit volume of the cell concentrate after cryoprotectant addition
Pump outlet pressure [min/max/avg]	Pump outlet	%	179/171	2.415 (0.066)[d]	Pump operating output expressed as a percentage of the maximum pump capacity during the concentration process
Flow rate [min/max/avg]	–	kL/h	193/178	2.438 (0.064)[d]	Volumetric flow rate of the culture broth measured during the concentration process
Process pressure [min/max/avg]	–	bar	193/178	2.444 (0.064)[d]	Internal pressure of the culture vessel during the centrifugation process
Back pressure [min/max/avg]	–	kPa	180/171	2.205 (0.072)[d]	Pressure generated by downstream flow resistance during the concentration process
Cryoprotectant volume[e]	CP volume	kg	196/178	–	Amount of cryoprotectant added to the cell concentrate prior to freeze-drying
Number of discharges	Discharge [n]	–	190/178	–	Total number of discharge events occurring during centrifugation
Average volume per discharge	Discharge [avg]	kg	190/178	–	Average mass of material discharged per centrifuge discharge event
Centrifugation time	–	h	190/178	–	Total duration of the centrifugation process
Freeze-dried cake[f]	–	kg	196/178	–	Mass of the freeze-dried product obtained after completion of freeze-drying
Total production[f]	–	kg	184/168	–	Total mass of the final product recovered after completion of all downstream processes

[a] Number of batches with a non-missing value among the raw production batches (n = 196) and the batches included after preprocessing (n = 178). One batch-level value was retained for each variable after within-batch measurements were summarized, where applicable.

[b] For variables measured repeatedly within a batch, the mean number of measurements and the standard error of the mean (SEM) are reported across the batches included after preprocessing (n = 178).

[c] TCC and VCC were measured in technical triplicate within each batch, and the mean of the three measurements was used as a single batch-level value.

[d] The four time-series process variables were recorded at one to four time points per batch.

[e] Excluded from the VCC reduction prediction.

[f] Excluded from both the total production and VCC reduction predictions. Variables without superscript e or f were used as inputs in all three prediction tasks.

informed by evidence that imputing variables with missing rates of 50% or higher yields results deviating substantially from completely observed data (Junaid et al., 2025). Step (ii) was based on the rationale that samples with excessive missingness retain little reliable information even after imputation (Oexner et al., 2025). Step (iii) applied a stricter threshold to exclude variables whose imputation quality could not be reliably ensured, given the relatively large number of variables compared to the number of samples (Abu-Mostafa et al., 2012).

After filtering, the final dataset comprised 178 batch samples with 36 variables: 34 process variables derived from measurements available prior to freeze-drying, freeze-dried cake weight, and total production. Descriptions and abbreviations are provided in Table 1 for the 36 variables, and in Supplementary Table S1 for the excluded variables. Two additional variables — VCC measured after post-processing and one week after sieving — were included to calculate quality attributes.

2.3.2. Missing value imputation

Remaining missing values were imputed using variable-specific strategies (Kraus and Bartsch, 2026): median values (Zhou et al., 2024; Little and Rubin, 2019) for volume- and inoculum-related variables; linear interpolation (Lepot et al., 2017) across batches ordered by the TCC at the corresponding or preceding stage, for pH-related variables and TCC in the culture broth; and mean values (Zhou et al., 2024; Little and Rubin, 2019) for all other variables. Samples lacking all centrifugation-related variables (number of discharges, average volume per discharge, and centrifugation time) were excluded, as their simultaneous absence rendered the samples uninformative for characterizing centrifugation behavior.

2.4. Predictive modeling and evaluation

To enable an exploratory assessment of relative batch-to-batch quality variability, classification models were developed to predict quality variability associated with three batch-level attributes: (1) total production, (2) VCC reduction relative to the theoretical cell count in the cell concentrate, and (3) change in VCC after one week relative to the post-processing VCC. These attributes represent complementary aspects of process performance, including production yield, immediate viable-cell loss during downstream processing, and post-process stability.

2.4.1. Definition of quality variables

Total production was defined as the total mass of the final product after completion of all downstream processes. The theoretical post-processing VCC was calculated using the operational equation provided by the manufacturer for the production dataset used in this study, which is based on a material balance of viable cells across downstream processing [Doran, 2013; Bérangère et al., 2012]:

$$\mathrm{VCC}_{\mathrm{theoretical}} = \frac{(\mathrm{volume}_{\mathrm{cc}} + \mathrm{CP\ volume}) \times \mathrm{VCC}_{\mathrm{cc}}}{\mathrm{total\ production}}, \quad (1)$$

where $\mathrm{volume}_{\mathrm{cc}}$ and $\mathrm{VCC}_{\mathrm{cc}}$ denote the recorded amount and VCC of the cell concentrate, respectively. The cell concentrate amount and CP volume were recorded in kg (Table 1). A density of 1 kg/L was assumed to convert these recorded masses to their corresponding volumes (Doran, 2013). The theoretical VCC was expressed as CFU/g after converting L to mL and kg to g. Because both conversions involve a factor of 1000, these factors cancel in the numerical calculation shown in Eq. (1). This formulation provides a normalized reference for evaluating cell loss during downstream processing. VCC reduction was defined as the difference between the theoretical VCC and the measured post-processing VCC (Bérangère et al., 2012; van der Linden et al., 2021):

$$\mathrm{VCC\ reduction} = \mathrm{VCC}_{\mathrm{theoretical}} - \mathrm{VCC}_{\mathrm{post\text{-}processing}}, \quad (2)$$

VCC change after one week was defined as:

$$\mathrm{VCC\ change} = \mathrm{VCC}_{\mathrm{after\ 1w}} - \mathrm{VCC}_{\mathrm{post\text{-}processing}}, \quad (3)$$

which represents the change in viable cell concentration during one week of post-processing storage (Berninger et al., 2018; Jannah et al., 2022).

The distributional characteristics of these variables are summarized in Table 2. Total production ranged from 116.60 to 350.00 kg, with a mean of 238.52 kg (±3.58 kg). VCC reduction (vs. theoretical) ranged from 239.85 to 16,775 $\times 10^8$ CFU/g, with a mean of 5713.22 $\times 10^8$ CFU/g (±205.06 $\times 10^8$ CFU/g), indicating substantial inter-batch variability in viable cell loss during downstream processing. VCC change after one week ranged from −7731 to 1342 $\times 10^8$ CFU/g, with a mean of −877.34 $\times 10^8$ CFU/g (±79.60 $\times 10^8$ CFU/g). Most values were negative, reflecting an overall decline in VCCs during storage, although the magnitude varied across samples.

2.4.2. Input and target variables

Continuous quality variables were converted into binary classification targets using the mean value of each variable as the cutoff (mean values are given in Table 2). The resulting below-mean and above-mean groups comprised 83 and 95 batches for total production, 101 and 77 batches for VCC reduction, and 84 and 94 batches for VCC change, respectively. These data-derived cutoffs were used solely to distinguish batches with relatively lower and higher values for each quality attribute and did not represent industrial pass/fail standards. Because dichotomizing a continuous variable discards information on within-group differences and makes the resulting classes dependent on the chosen cutoff (MacCallum et al., 2002), the robustness of this choice was examined using an alternative median cutoff (Section 2.5.2).

For each quality variable, we first defined the intended prediction time in an actual industrial setting. Only variables that would be available at that time and were not directly used to calculate the corresponding quality variable were included as input features. These criteria were applied to prevent potential data leakage (Kapoor and Narayanan, 2023). The prediction time for total production was defined as the point after CP addition but before freeze-drying began. The prediction time for VCC reduction was defined as an earlier point, after completion of the concentration process but before CP addition. The prediction time for VCC change was defined as the point after post-processing but before the one-week re-analysis results became available.

Variable availability at a given prediction time also depends on the analytical method used. Under ILDONG Bioscience's current testing protocol, TCC is determined by direct hemocytometer counting and is generally available on the same day as sampling. In contrast, VCC requires MRS agar pour plating (Davis, 2014) followed by 48–72 h of anaerobic incubation at 37 °C, so results become available only 2–3 days after sampling. Accordingly, VCC measured at the main culture and cell concentrate stages, despite being sampled earlier in the process, was treated as unavailable at the prediction times for total production and VCC reduction, given the required incubation period.

Applying these criteria, the input variables for predicting total production consisted of 32 of the 34 pre-freeze-drying variables, excluding VCC of the main culture and cell concentrate. Freeze-dried cake weight and total production were measured after the prediction time and were therefore not included among these 34 candidate variables. For VCC reduction, CP volume and cell concentrate volume were additionally excluded because both variables are directly used to calculate theoretical VCC (Eq. (1)). CP volume was also unavailable at the intended prediction time because cryoprotectant had not yet been added. Cell concentrate VCC and total production, the other variables used in Eq. (1), had already been excluded based on prediction timing. Therefore, none of the variables used to calculate theoretical VCC were included as inputs, resulting in 30 input variables for VCC reduction. For VCC change, all 36 variables listed in Table 1, including VCC of the main culture and cell concentrate, were available at the prediction time and were used as inputs. Table 1 summarizes the input composition for each prediction task.

Table 2
Descriptive statistics of the quality variables used in this study.

Statistic	Quality variable		
	Total production (kg)	VCC reduction (vs. theoretical, $\times 10^8$ CFU/g)	VCC change (after 1w, $\times 10^8$ CFU/g)
[Min, Max]	[116.60, 350.00]	[239.85, 16,775.00]	[−7731.00, 1342.00]
Mean (SEM)	238.52 (3.58)	5713.22 (205.06)	−877.34 (79.60)
Median [Q1, Q3]	250.00 [200.00, 300.00]	5172.81 [3876.88, 7407.71]	−823.50 [−1457.75, −142.50]

SEM: standard error of the mean; Q1: 25th percentile; Q3: 75th percentile.

Table 3
Classification performance of the XGBoost models for each quality variable.

Quality variable	Metrics (%), mean (SEM)			
	Accuracy	F1	Precision	Sensitivity
Total production	75.29 (2.42)	75.29 (3.28)	78.17 (3.47)	74.41 (4.82)
VCC reduction	74.25 (3.17)	67.47 (5.86)	71.44 (7.32)	66.29 (6.47)
VCC change	59.48 (3.81)	59.79 (4.67)	62.87 (5.75)	60.27 (6.15)

The mean scores and standard errors of the mean (SEM) across 10 folds were reported.

2.4.3. Classification scheme and modeling

XGBoost was employed due to its ability to capture nonlinear relationships and handle the heterogeneous process variables commonly observed in industrial bioprocess data (Chen and Guestrin, 2016; Paul et al., 2025; Ageh et al., 2024; Saha et al., 2025). The model parameter settings are summarized in Supplementary Table S2. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002; Lemaître et al., 2017) was applied exclusively to the training data within each cross-validation fold (Kapoor and Narayanan, 2023; Kuhn and Johnson, 2019; Vandewiele et al., 2021). Median and mean imputation of missing values (Section 2.3.2) was also performed within each cross-validation fold. Imputation values were computed using only the training-fold data and then applied to the corresponding held-out fold (Kapoor and Narayanan, 2023; Kuhn and Johnson, 2019). Given the limited number of available batches, no separate external test set was used. Model performance was evaluated using 10-fold cross-validation (Kohavi, 1995), in which the dataset was randomly partitioned into 10 approximately equal folds. In each of the 10 iterations, nine folds were used for model training, while the remaining fold was held out exclusively for performance evaluation (Nti et al., 2021). Each batch was included in the held-out evaluation fold exactly once. Four performance metrics were utilized (Sokolova and Lapalme, 2009): accuracy, F1-score, precision, and sensitivity (recall).

2.4.4. Feature importance analysis

To further interrogate the predictive contributions of individual feature variables, feature importance scores derived from the trained XGBoost models were systematically analyzed (Shi et al., 2019). In tree-based ensemble models such as XGBoost, feature importance reflects the contribution of each variable to the reduction of the loss function through recursive partitioning, thereby capturing nonlinear relationships and interactions among variables (Chen and Guestrin, 2016; Ageh et al., 2024). To obtain a stable and representative estimate, the importance values were averaged across the ten models trained during cross-validation (Molnar et al., 2022). Relative importance was reported, which is defined as the proportion of each variable to the total importance across all features, in the following results section.

2.5. Sensitivity and comparative analyses

2.5.1. Missing-rate threshold sensitivity analysis

The main analysis applies a three-step missing-rate thresholding procedure for handling missing values (see Section 2.3.1). However, the sample-level threshold of 25.0% in step (ii) and the variable-level threshold of 25.0% in step (iii) were chosen empirically to ensure overall data completeness and are thus somewhat arbitrary. Because such data-processing decisions can influence downstream results (Steegen et al., 2016), we conducted a sensitivity analysis (Thabane et al., 2013) using four alternative threshold combinations: fixing the variable-level threshold (25.0%) while setting the sample-level threshold to 12.5% (stricter) or 50.0% (relaxed), and conversely, fixing the sample-level threshold (25.0%) while setting the variable-level threshold to 12.5% or 50.0%. For each resulting dataset, imputation was performed following the same procedure as the main analysis, and the same classification scheme and feature analysis protocol were applied for comparison.

2.5.2. Classification cutoff sensitivity analysis

To convert relative batch-to-batch quality variation into a binary classification problem, the main analysis used the mean of each quality variable as the cutoff to divide the 178 batches into below- and above-mean groups (see Section 2.4.2). Because group membership near a data-derived cutoff can shift with the cutoff definition (MacCallum et al., 2002), the median was used as an alternative cutoff (median values are given in Table 2) to define below- and above-median groups for each quality variable: total production into above-median (n = 93) and below-median (n = 85), VCC reduction into above-median (n = 90) and below-median (n = 88), and VCC change into below-median (n = 92) and above-median (n = 86). The same predictive modeling and evaluation protocol as in the main analysis was then applied to the reclassified dataset for comparison.

2.5.3. Baseline model comparison

To evaluate the predictive performance of XGBoost relative to models based on different learning principles, we additionally trained support vector machine (SVM; margin-based) (Onel et al., 2018; Cortes and Vapnik, 1995) and random forest (bagging-based ensemble) (Melcher et al., 2015; Breiman, 2001) models. We compared their classification performance to assess whether the observed predictive performance depended substantially on the choice of learning algorithm. The models were implemented using the xgboost (v3.0.2) (Chen and Guestrin, 2016), scikit-learn (v1.7.0) (Pedregosa et al., 2011), and imbalanced-learn (v0.14.1) (Lemaître et al., 2017) Python packages, and the parameter settings are summarized in Supplementary Table S2.

3. Results

3.1. Dataset characteristics and quality attributes

Descriptive statistics of key process/quality variables stratified by classification group are presented in Supplementary Table S3. At the overall level, process-scale variables such as main culture volume (approximately 20,500 L) and cell concentrate mass (approximately 617 kg) showed relatively limited variability. In contrast, differences between groups were observed in variables associated with downstream processing and cell viability. In particular, the above-mean total production group exhibited higher values for both cell concentrate volume and CP volume compared to the below-mean, suggesting a potential association between biomass availability, formulation conditions, and final production yield.

Table 4
Relative feature importance of top variables contributing to approximately 50% of the total importance for predicting each quality variable using the XGBoost model.

Rank	Total production		VCC reduction (vs. theoretical)		VCC change (after 1w)	
	Variable	Importance (%)	Variable	Importance (%)	Variable	Importance (%)
1	CP volume	14.52	Back pressure [avg]	10.52	CP volume	10.60
2	Process pressure [min]	11.74	Discharge [n]	10.38	Total production	8.92
3	Pump outlet [avg]	7.47	Process pressure [min]	8.17	Discharge [n]	7.37
4	Discharge [n]	6.99	Pre-inoculation pH	7.05	Volume (cell concentrate)	5.08
5	Discharge [avg]	5.72	TCC (1st seed)	6.41	VCC (cell concentrate)	4.95
6	TCC (2nd seed)	5.61	Final pH (2nd seed)	5.66	Final pH (2nd seed)	4.95
7			Flow rate [avg]	5.01	Volume (pre-culture)	4.41
8					Freeze-dried cake	4.38
Total		52.06		53.20		50.67

Table 5
Descriptive statistics for the top variables in the importance analysis of a stratified dataset by classification group.

Total production			VCC reduction (vs. theoretical)			VCC change (after 1w)		
	Below-mean	Above-mean		Below-mean	Above-mean		Below-mean	Above-mean
CP volume	207.00 (2.84)	248.74 (8.84)	TCC (1st seed)	59.45 (1.32)	56.24 (1.11)	Volume (pre-culture)	7256.41 (200.26)	7347.83 (190.83)
TCC (2nd seed)	53.98 (1.14)	51.93 (1.23)	Final pH (2nd seed)	4.54 (0.02)	4.61 (0.02)	Final pH (2nd seed)	4.53 (0.02)	4.60 (0.02)
Process pressure [min]	0.58 (0.04)	0.82 (0.03)	Pre-inoculation pH	4.52 (0.02)	4.60 (0.03)	Discharge [n]	44.64 (1.07)	46.07 (1.18)
Pump outlet [avg]	22.71 (1.23)	19.47 (0.97)	Process pressure [min]	0.78 (0.03)	0.60 (0.04)	Volume (cell concentrate)	616.04 (9.22)	618.31 (9.82)
Discharge [n]	41.27 (1.14)	49.00 (0.99)	Flow rate [avg]	7.16 (0.15)	7.63 (0.09)	VCC (cell concentrate)	3385.37 (71.69)	3189.72 (62.18)
Discharge [avg]	16.05 (0.65)	13.49 (0.36)	Back pressure [avg]	220.79 (0.68)	221.13 (0.46)	CP volume	222.68 (9.13)	235.17 (5.26)
			Discharge [n]	48.25 (1.01)	41.64 (1.17)	Freeze-dried cake	180.41 (2.65)	185.96 (3.19)
						Total production	236.11 (5.02)	240.67 (5.09)

For each prediction variable, the mean and standard error of the mean (SEM) calculated from the samples in each classification group were reported—total production: below-mean ($n = 83$), above-mean ($n = 95$); VCC reduction: below-mean ($n = 101$), above-mean ($n = 77$); VCC change: below-mean ($n = 84$), above-mean ($n = 94$).

3.2. Prediction performance for quality variability

The classification performance of the XGBoost models for the three quality variables is summarized in Table 3, and confusion matrices are presented in Supplementary Fig. S1, illustrating the classification tendencies underlying the performance metrics reported in Table 3.

The models achieved comparable performance for total production and VCC reduction, with mean accuracies of 75.29% (SEM: ±2.42%) and 74.25% (±3.17%), respectively. For total production, precision was 78.17% (±3.47%) and sensitivity was 74.41% (±4.82%), indicating broadly comparable performance across the two metrics. For VCC reduction, precision (71.44%, ±7.32%) and sensitivity (66.29%, ±6.47%) were both moderately lower than the corresponding values for total production. In contrast, the prediction of VCC change during one-week post-processing storage showed the lowest performance among the three quality variables, with an accuracy of 59.48% (±3.81%), F1-score of 59.79% (±4.67%), precision of 62.87% (±5.75%), and sensitivity of 60.27% (±6.15%).

3.3. Key process factors associated with quality variability

To facilitate interpretable analysis, variables accounting for approximately 50% of the cumulative feature importance were selected (Table 4). Descriptive statistics of the top variables, stratified by classification group (below- and above-mean), are presented in Table 5.

For total production, six variables accounted for 52.06% of the total importance, with CP volume as the highest-ranked predictor (14.52%), followed by minimum process pressure (11.74%) and mean pump outlet pressure (7.47%). Discharge-related variables, including the number of discharges (6.99%) and average volume per discharge (5.72%), and second seed culture TCC (5.61%) were also included. Batches with above-mean total production exhibited a higher mean CP volume (248.74 vs. 207.00; Table 5) and a higher mean number of discharges (49.00 vs. 41.27) relative to below-mean batches, whereas mean pump outlet pressure was lower in the above-mean group (19.47 vs. 22.71).

For VCC reduction, seven variables explained 53.20% of the total importance, led by mean back pressure (10.52%) and the number of discharges (10.38%), followed by minimum process pressure (8.17%), pre-inoculation pH (7.05%), first seed culture TCC (6.41%), second seed culture final pH (5.66%), and mean flow rate (5.01%). Meanwhile, the group-wise comparison of mean back pressure, which had the highest importance, showed nearly zero difference between below-mean and above-mean batches (Table 5).

For VCC change, eight variables were required to account for 50.67% of the total importance, reflecting a more distributed contribution across features. CP volume (10.60%) and total production (8.92%) were the two leading contributors, followed by the number of discharges (7.37%), cell concentrate volume (5.08%), cell concentrate VCC (4.95%), second seed culture final pH (4.95%), pre-culture volume (4.41%), and freeze-dried cake weight (4.38%).

3.4. Effects of missing-data thresholds

To examine the effect of the missing-rate-based preprocessing thresholds, four alternative threshold combinations were additionally examined. When the sample-level threshold in step (ii) was fixed at 25% and only the variable-level threshold in step (iii) was varied, the number of input samples remained unchanged at 178, and the number of input variables increased as the threshold was relaxed (total production: 28 to 35, VCC reduction: 26 to 33, VCC change: 32 to 39 for 12.5% to 50.0%), and no additional variables were excluded in step (iii) when applying the 50% threshold. When the variable-level threshold was fixed at 25%, and the sample-level threshold was varied, the number of samples decreased to 151 under the stricter threshold (12.5%) and increased to 190 under the more relaxed threshold (50%).

In terms of classification performance (Table 6), the prediction of total production remained relatively stable across variations in the variable-level threshold (mean accuracy: 74.74%–75.29%). However, it was considerably affected by changes in the sample-level threshold: mean accuracy varied from 69.54% (stricter threshold) to 77.37% (relaxed threshold). The prediction of VCC reduction was comparatively sensitive to changes in the variable-level threshold, with mean accuracy decreasing from 74.25% in the main analysis (25%) to 68.04% and 70.82% under the thresholds 12.5% and 50.0%, respectively. A similar pattern was observed for the sample-level threshold, with reduced performance under both variants: mean accuracy 65.50% and 63.16% under the thresholds 12.5% and 50.0%, respectively, indicating a modest degree of sensitivity to threshold changes in either direction. The prediction of VCC change was the most sensitive to changes in

Table 6
Comparison of preprocessed dataset configurations and classification performance in the missing-rate threshold sensitivity analysis.

Missing rate threshold	Model input		Metrics (%), mean (SEM)		Missing rate threshold	Model input		Metrics (%), mean (SEM)	
sample/variable[a]	# samples	# variables	Accuracy	F1	sample/variable	# samples	# variables	Accuracy	F1
25.0%/12.5%					12.5%/25.0%				
Total production	178	28	**75.29 (2.20)**	75.26 (2.80)	Total production	151	33	69.54 (2.65)	69.69 (2.54)
VCC reduction (vs. theoretical)	178	26	68.04 (2.11)	62.05 (3.89)	VCC reduction (vs. theoretical)	151	31	65.50 (2.83)	56.06 (5.24)
VCC change (after 1w)	178	32	50.03 (3.50)	51.88 (4.07)	VCC change (after 1w)	151	37	43.08 (5.31)	46.17 (5.74)
25.0%/25.0%					25.0%/25.0%				
Total production	178	32	**75.29 (2.42)**	**75.29 (3.28)**	Total production	178	32	75.29 (2.42)	75.29 (3.28)
VCC reduction (vs. theoretical)	178	30	**74.25 (3.17)**	**67.47 (5.86)**	VCC reduction (vs. theoretical)	178	30	**74.25 (3.17)**	**67.47 (5.86)**
VCC change (after 1w)	178	36	**59.48 (3.81)**	**59.79 (4.67)**	VCC change (after 1w)	178	36	**59.48 (3.81)**	**59.79 (4.67)**
25.0%/50.0%[b]					50.0%/25.0%				
Total production	178	35	74.74 (2.69)	75.03 (3.18)	Total production	190	31	**77.37 (2.49)**	**77.63 (3.12)**
VCC reduction (vs. theoretical)	178	33	70.82 (2.42)	63.61 (5.45)	VCC reduction (vs. theoretical)	190	29	63.16 (3.92)	58.28 (5.57)
VCC change (after 1w)	178	39	52.22 (3.53)	55.29 (4.10)	VCC change (after 1w)	190	35	47.37 (3.68)	48.21 (4.55)

[a] Sample-level threshold and variable-level threshold used at the second and third step of missing-rate-based exclusion, respectively.

[b] Applying a 50% variable-level threshold did not exclude any variables. The underlined threshold pairs represent the values used in the main analysis. This shows the classification performance of the XGBoost model along with the number of input samples and the number of input variables, and displays the mean score across 10 folds and the standard error of the mean (SEM). The left side of the table shows the comparison when the sample-level threshold is fixed, while the right side shows when the variable-level threshold is fixed. The bold indicates the best mean score for each comparison and each target quality variable.

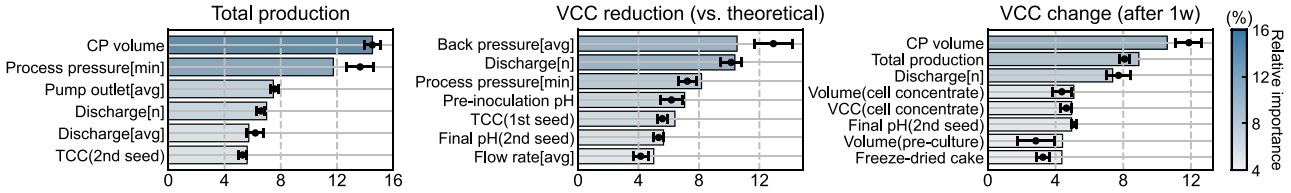


Fig. 2. Feature importance of threshold sensitivity analysis. Each bar represents the importance (%) in the main analysis; each error bar shows the standard error of the mean (SEM) for the five missing-rate threshold pairs, and the mean scores are indicated by markers.

Table 7
Comparison of classification performance for cutoff sensitivity analysis.

	Metrics (%), mean (SEM)			
	Accuracy	F1	Precision	Sensitivity
Median cutoff				
Total production	70.82 (2.45)	70.57 (3.86)	72.29 (4.74)	70.59 (4.87)
VCC reduction	71.44 (2.73)	**70.34 (3.75)**	71.26 (4.79)	**70.96 (4.17)**
VCC change	53.30 (2.50)	51.36 (3.88)	53.49 (4.72)	51.21 (4.74)
Mean cutoff				
Total production	**75.29 (2.42)**	**75.29 (3.28)**	**78.17 (3.47)**	**74.41 (4.82)**
VCC reduction	**74.25 (3.17)**	67.47 (5.86)	**71.44 (7.32)**	66.29 (6.47)
VCC change	**59.48 (3.81)**	**59.79 (4.67)**	**62.87 (5.75)**	**60.27 (6.15)**

The classification performance of the XGBoost model was reported as the mean score and standard error of the mean (SEM) across 10 folds. Underlining indicates the cutoff used in the main analysis. The bold indicates the highest mean scores for each performance metric and each target quality variable.

the preprocessing conditions. Mean accuracy ranged from 50.03% to 59.48% across variations in the variable-level threshold, and the range was even wider for the sample-level threshold, varying from 43.08% (stricter threshold) to 59.48% (main analysis).

In the feature importance analysis (Fig. 2 and Supplementary Table S4), for total production, the importance of top-ranked variables, including CP volume and minimum process pressure, remained stable across the five threshold combinations, exhibiting relatively narrow error ranges (SEM) and values comparable to or slightly higher than those of the main analysis. For VCC reduction, although the top-ranked variable, mean back pressure, showed a relatively wide error range, this range was shifted to higher values than in the main analysis, such that the ranking of variables was not affected. For VCC change, most top-ranked variables, including CP volume and number of discharges, exhibited error ranges that were considerably wide relative to the bar heights, while the relatively lower-ranked variables were distributed at lower values, indicating that the importance of lower-ranked variables could vary substantially depending on the threshold condition.

3.5. Effects of the classification cutoff

Quality classes were redefined using the median rather than the mean cutoff used in the main analysis. The XGBoost models were then retrained using the redefined classes, and their performance was compared with that of the main analysis (Table 7). For total production and VCC change, all four performance metrics decreased under the median cutoff. Mean accuracy decreased from 75.29% to 70.82% for total production and from 59.48% to 53.30% for VCC change, with the larger decline observed for VCC change. For VCC reduction, mean accuracy decreased slightly from 74.25% to 71.44%, whereas F1-score and sensitivity increased under the median cutoff.

In the feature importance analysis (Fig. 3, Supplementary Table S5), the prediction of total production showed the most stable pattern, with the six variables accounting for the top 50% of importance (CP volume, minimum process pressure, and others) remaining completely unchanged from the main analysis. For VCC reduction, the number of top-ranked variables decreased from seven to four, but mean back pressure and number of discharges remained the top-ranked variables. For VCC change, five variables — number of discharges, CP volume, total production, second seed culture final pH, and cell concentrate VCC — were commonly retained from the main analysis, whereas the remaining variable composition differed to some extent.

3.6. Performance comparison with baseline models

Table 8 summarizes the predictive performance of baseline models (SVM and random forest) and XGBoost. For total production,

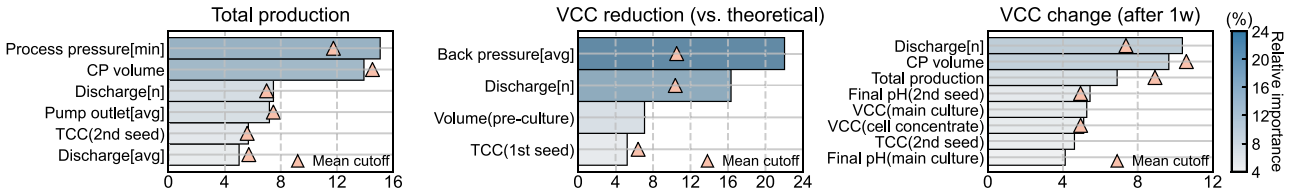


Fig. 3. Comparison of feature importance from classification cutoff sensitivity analysis. Each panel shows the relative feature importance (%) of top variables contributing to approximately 50% in each quality variable prediction; the bars represent the importance from the median cutoff, and the triangle markers indicate the importance from the mean cutoff (main analysis) if the variable is included within the top variables of the main analysis (Table 4).

Table 8
Comparison of classification performance with baseline models.

	Metrics (%), mean (SEM)			
	Accuracy	F1	Precision	Sensitivity
SVM				
Total production	74.31 (4.59)	**76.97 (4.43)**	74.02 (5.48)	**81.43 (4.15)**
VCC reduction	71.41 (3.11)	66.97 (4.56)	65.49 (4.64)	**69.52 (5.28)**
VCC change	58.43 (3.53)	59.33 (3.55)	**63.69 (6.07)**	58.95 (4.55)
Random forest				
Total production	72.55 (2.74)	73.31 (3.66)	74.33 (4.48)	74.60 (5.24)
VCC reduction	72.55 (3.63)	66.61 (6.00)	68.18 (6.47)	67.30 (7.02)
VCC change	56.21 (6.59)	59.71 (5.81)	63.34 (8.75)	**60.99 (5.79)**
XGBoost				
Total production	**75.29 (2.42)**	75.29 (3.28)	**78.17 (3.47)**	74.41 (4.82)
VCC reduction	**74.25 (3.17)**	**67.47 (5.86)**	**71.44 (7.32)**	66.29 (6.47)
VCC change	**59.48 (3.81)**	**59.79 (4.67)**	62.87 (5.75)	60.27 (6.15)

The mean scores and standard errors of the mean (SEM) across 10 folds were reported. The underline indicates the model (XGBoost) used in the main method. The bold indicates the highest mean scores for each metric and each target quality variable across the models.

mean accuracy was similar across the three models (72.55%–75.29%), with XGBoost achieving the highest accuracy (75.29%) and precision (78.17%), while SVM showed the highest F1-score (76.97%) and sensitivity (81.43%). For VCC reduction, XGBoost achieved the best mean accuracy (74.25%) and F1-score (67.47%), though differences among models were modest (mean accuracy: 71.41%–74.25%; mean F1: 66.61%–67.47%). VCC change showed the lowest performance overall (mean accuracy: 56.21%–59.48%); XGBoost showed a slight edge in accuracy and F1-score, but SVM had higher precision (63.69%), indicating no clear superiority of any single model. The comparable performance across the three models indicates that the observed results were not unique to XGBoost. However, this within-dataset comparison does not replace validation using an independent external dataset.

4. Discussion

This study trained XGBoost models for three quality variables — total production, VCC reduction at post-processing relative to theoretical value, and VCC change after one week — to evaluate classification performance and analyze feature importance. Total production refers to the weight of the final product recovered after milling and sieving; VCC reduction refers to the extent to which the viable cell count per unit mass decreased relative to the theoretical value, i.e., immediate cell survival following processing; and VCC change refers to the change in viable cell count per unit mass after one week of storage, i.e., cell survival after storage for a defined period. Our results confirmed that the predictions of VCC reduction and VCC change should be interpreted cautiously, focusing on a subset of high-importance variables.

In terms of classification performance (Table 3), total production achieved the highest performance, VCC reduction showed relatively good performance, and VCC change showed comparatively lower performance. In the feature importance analysis (Table 4), CP volume and minimum process pressure each accounted for more than 10% importance for total production, with mean pump outlet pressure and number of discharges at approximately 7%. For VCC reduction, mean back pressure and number of discharges accounted for more than 10%, and minimum process pressure and pre-inoculation pH for more than 7%. For VCC change, CP volume accounted for more than 10% and number of discharges for more than 7%, with variables directly related to the theoretical cell count calculation (total production, cell concentrate volume/VCC) also showing high importance. Unlike the other targets, importance for VCC change was distributed relatively evenly across variables, suggesting that predicting cell survival following storage may require more complex, multivariate relationships or that the measured variables are insufficient to fully explain it—consistent with its lower classification performance. The results for VCC change should therefore be regarded as an exploratory and suggestive trend rather than a definitive conclusion.

A sensitivity analysis examining the effect of missing-rate thresholds showed that total production remained relatively stable, VCC reduction was particularly sensitive to the variable-level threshold, and VCC change was the most sensitive overall (Table 6). In the feature importance analysis, top-ranked variables for total production remained comparable to or higher than in the main analysis (Fig. 2), whereas for VCC reduction and especially VCC change, interpretation should focus on the subset of top-ranked variables that maintained low sensitivity.

The median cutoff analysis, conducted to verify robustness, supported this trend: performance variation was largest for VCC change (Table 7), and while the top-50% variable composition for total production remained identical to that of the mean cutoff, only a subset of the highest-ranked variables was retained for VCC reduction and VCC change (Supplementary Table S5). Accordingly, results were interpreted with a focus on variables showing stable top-ranked importance across these robustness analyses: for total production, CP volume, centrifugation-related variables (number of discharges, average volume per discharge), and pressure-related variables (minimum process pressure, mean pump outlet pressure); for VCC reduction, number of discharges and mean back pressure; and for VCC change, CP volume and number of discharges. Because the avg/min/max variables derived from each time-series process variable (pump outlet pressure, flow rate,

process pressure, back pressure) exhibit high mutual multicollinearity, the specific variable selected among collinear sets can vary with minor data changes or retraining (Kim, 2019). Interpretation therefore focused on the corresponding upstream process variable rather than the individual avg/min/max statistics.

Notably, variables related to the concentration and freeze-drying processes showed higher importance and greater stability than variables directly related to production quality. TCC-related variables were included as inputs for total production and VCC reduction, and VCC change additionally included total production- and VCC-related variables (Table 1). In the main analysis, seed culture TCC variables were among the top 50% importance variables for total production and VCC reduction but contributed less than 6% (Table 4); similarly, variables used to calculate the theoretical post-processing VCC (total production, cell concentrate volume/VCC; Eq. (1)) were among the top 50% for VCC change but showed lower importance than CP volume and number of discharges, or high variability in the additional analyses. Interpretation therefore focused on the concentration process (pressure- and discharge-related variables) and the freeze-drying process (CP volume).

CP volume showed particularly prominent importance for total production and VCC change, with higher values observed in the above-mean group for both (Supplementary Table S3). As this does not guarantee a linear relationship, several interpretations should be considered. First, an appropriate CP volume may have reduced losses during milling/sieving: freeze-dried cake inevitably loses material during this process (Oyinloye and Yoon, 2020), and the strength and uniformity of its porous structure depend strongly on the solids concentration of the initial slurry, i.e., the dilution determined by CP volume (Kasper and Friess, 2011; Hedberg et al., 2019). An appropriate CP volume may thus allow the cake to fracture consistently into powder matching the target particle size distribution, increasing sieve pass-through and production yield (Touzet et al., 2018).

Second, CP volume may have acted through its intrinsic cryoprotectant role, protecting cells from freezing and drying stress (Ge et al., 2024). A higher CP volume, given similar total input cell counts, may indicate a sufficient final cryoprotectant concentration, reducing ice-crystal-induced membrane damage (Li et al., 2025), enhancing vitrification-based protection of protein/membrane structure during drying (Noufeu et al., 2025), and mitigating gradual storage deterioration from residual moisture and reactive oxygen species (Nag and Das, 2013)—potentially explaining its importance for VCC change. Although CP volume was not included as an input for VCC reduction prediction, if this hypothesis holds, a higher CP volume could jointly improve survival immediately after freeze-drying and reduce further cell death during storage.

However, CP volume may also have been merely a secondary indicator of underlying process status: batches with more favorable culture outcomes may simply have received more CP volume to reach the target VCC concentration, reflecting an inherently favorable batch rather than an effect of CP volume itself. As these results do not guarantee linearity or causality, this possibility should be considered.

The number of discharges showed high importance across all predictions, and average volume per discharge was also top-ranked for total production, indicating that discharge-related variables are relevant to both recovery and short- and long-term survival. Group-wise comparisons (Table 5) showed more frequent discharge in batches with higher total production, lower VCC reduction, and larger VCC change. For total production, the lower average volume per discharge in the higher-recovery group suggests shortened discharge intervals rather than simply a larger culture yield. Prolonged discharge intervals may thicken the cake inside the centrifuge bowl, reducing separation efficiency and allowing cells to escape into the effluent (Baust et al., 2025), while prolonged retention in the cake may expose cells to pressure, friction, and heat, increasing cell death or impairing long-term stability (Zaman et al., 2009; Peterson et al., 2012; Katagiri et al., 2022). As with CP volume, however, this pattern may instead reflect an inherently favorable, high-solids-content batch in which solids accumulated more rapidly, naturally leading to more frequent, shorter discharges (Richardson and Walker, 2018), representing the same type of confounding possibility; these results likewise do not guarantee linearity or causality.

Specifically for total production, process pressure and pump outlet pressure showed high importance, with higher process pressure and lower pump outlet pressure in the higher-recovery group (Table 5). As both are measured upstream of centrifugation, a higher process pressure may strengthen the transfer driving force, allowing lower pump load (Doran, 2013) and reflecting stable transfer flow—suggesting these variables indicate low overall process loss rather than directly affecting yield.

For VCC reduction, back pressure showed high importance despite a minimal mean difference between groups (Table 5). As XGBoost can capture complex nonlinear interactions (Chen and Guestrin, 2016), this may reflect interaction effects—for example, high back pressure with a short discharge interval (brief exposure) may have little effect on survival, whereas low back pressure with a long interval (prolonged exposure) may accumulate cellular stress. However, this remains an illustrative hypothesis. The high importance assigned to back pressure may also reflect model-dependent feature attribution or incidental patterns in this limited dataset rather than a stable relationship (Lundberg and Lee, 2017; Strobl et al., 2007). The nonlinear explanation presented above should therefore be regarded as a hypothesis requiring further validation.

The VCC change model, constructed to examine post-processing storage survival, showed low predictive performance and stability, warranting particular caution. The high importance of CP volume may support a hypothesis that reduced cellular damage during freeze-drying contributed to stable storage survival (Li et al., 2025; Noufeu et al., 2025), and the high importance of number of discharges may suggest that shorter exposure to discharge pressure was also relevant (Zaman et al., 2009); however, the model's failure to identify pressure-related variables as important suggests insufficient evidence for this hypothesis, possibly related to its low classification performance. Other unmeasured factors — such as storage temperature, packaging method, and oxygen exposure (Jannah et al., 2022) — may also influence long-term survival, suggesting that the data coverage in this study may not have been sufficient.

This study has several limitations. The analysis was based on data collected from a single industrial process, and no independent external test set was available. Model performance was therefore evaluated using cross-validation within this dataset. These constraints may limit the generalizability of the reported performance and findings to other strains, production systems, or operating conditions. In the absence of an industrially agreed quality standard, a data-driven cutoff (the mean value) was used to define the quality classes, and sensitivity analyses of the cutoff definition (mean vs. median) and preprocessing conditions confirmed that VCC reduction and VCC change predictions were somewhat sensitive to these methodological choices. In addition, it cannot be ruled out that these variables function as indirect proxies for overall batch favorability rather than as genuine causal factors, and the observational nature of the data precludes a definitive causal interpretation. Given the limited sample size of 178 batches, the feature importance derived from machine learning models may also reflect incidental patterns rather than a genuine signal, warranting caution in interpreting individual variable importance. Moreover, long-term outcomes such as VCC change may be influenced by unmeasured factors, including storage temperature, packaging method, and oxygen exposure.

Future studies should incorporate such storage-related variables and real-time monitoring data, validate generalizability using larger-scale, multi-site datasets, and clarify process-quality relationships through controlled experimental validation. Such efforts could develop the present approach into a predictive monitoring framework applicable in industrial settings.

5. Conclusion

The practical value of the present findings lies in narrowing the range of formulation and downstream processing conditions that warrant further investigation. Cryoprotectant volume, discharge characteristics, and pressure-related variables were repeatedly associated with variability in production yield or viable cell concentration. These variables should be regarded as candidates for process evaluation rather than established causal control factors. Several limitations define the scope of these findings. The analysis was based on 178 batches from a single manufacturing site, which limits its generalizability to other strains, facilities, and operating conditions. The classification targets were defined using data-derived mean cutoffs rather than predefined industrial quality criteria. In addition, the observational design cannot distinguish direct process effects from indirect associations or unmeasured confounding. The findings for one-week VCC change require particular caution because of their lower predictive performance, sensitivity to the cutoff definition, and the absence of detailed storage-related variables. Future studies should evaluate the identified associations in larger, multi-site manufacturing datasets and incorporate predefined quality specifications where available. Additional storage conditions should be recorded to improve the analysis of post-processing viability. Controlled process experiments will ultimately be required to determine whether modifying the identified variables produces reproducible improvements in product quality.

CRediT authorship contribution statement

Yunjeong Jang: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **BoYoung Choi:** Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation. **Roneel V. Sharan:** Writing – review & editing, Visualization, Validation. **Haeseong Park:** Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation. **Yohwan Kim:** Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation. **Jin Seok Moon:** Writing – review & editing, Visualization, Validation, Resources, Project administration, Methodology, Conceptualization. **Minji Lee:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Minji Lee reports financial support was provided by Ildong Bioscience Co., Ltd. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00336880; RS-2025-23963510). This study was also supported by Ildong Bioscience Co., Ltd., which provided research funding and study materials.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jfoodeng.2026.113386.

Data availability

The data that has been used is confidential.

References

Abdul Manan, Musaalbakri, 2025. Progress in probiotic science: Prospects of functional probiotic-based foods and beverages. Int. J. Food Sci. 2025 (1), 5567567.

Abu-Mostafa, Yaser S., Magdon-Ismail, Malik, Lin, Hsuan-Tien, 2012. Learning from Data: A Short Course. AMLBook.com.

Adli, Danung Nur, Fatyanosa, Tirana Noor, AlHuda, Fais, Sholikin, Mohammad Miftakhus, Sugiharto, Sugiharto, 2025. Modelling the growth performance and thermal environment of broiler chicken houses via different machine learning algorithms assisted by a customized Internet of Things. Smart Agric. Technol. 101421.

Afsharnia, F., Marzban, A., Abdeshahi, A., 2018. Sugarcane transportation process modeling by time series approach. Iran Agric. Res. 37 (2), 117–126.

Ageh, Opeyemi Micheal, Dasore, Abhishek, Hashim, Norhashila, Shamsudin, Rosnah, Man, Hasfalina Che, Ali, Maimunah Mohd, 2024. Quality monitoring of glutinous rice processing from drying to extended storage using hyperspectral imaging. Comput. Electron. Agric. 225, 109348.

Al Hagbani, Turki, Alamoudi, Jawaher Abdullah, Bajaber, Majed A., Alsayed, Huda Ibrahim, Al-Fanhrawi, Halah Jawad, 2025. Theoretical investigations on analysis and optimization of freeze drying of pharmaceutical powder using machine learning modeling of temperature distribution. Sci. Rep. 15 (1), 948.

Baust, Helene Katharina, Lam, Zihim, Hay, Michael, Nirschl, Hermann, Gleiß, Marco, 2025. Development of a dynamic process model for the mechanical fluid separation in disk stack centrifuges. Sep. Purif. Technol. 377, 134230.

Bérangère, Farges, Laurent, Poughon, Agnès, Pons, Claude-Gilles, Dussap, 2012. Methodology for bioprocess analysis: Mass balances, yields and stoichiometries. In: Innocenti, Alessio (Ed.), Stoichiometry and Research - The Importance of Quantity in Biomedicine. IntechOpen, London, chapter 12.

Berninger, Teresa, González López, Óscar, Bejarano, Ana, Preininger, Claudia, Sessitsch, Angela, 2018. Maintenance and assessment of cell viability in formulation of non-sporulating bacterial inoculants. Microb. Biotechnol. 11 (2), 277–301.

Box, George E.P., Jenkins, Gwilym M., Reinsel, Gregory C., Ljung, Greta M., 2015. Time Series Analysis: Forecasting and Control. John Wiley & Sons.

Breig, Sura Jasem Mohammed, Luti, Khalid Jaber Kadhum, 2021. Response surface methodology: A review on its applications and challenges in microbial cultures. Mater. Today Proc. 42 (5), 2277–2284.

Breiman, Leo, 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Buahom, Juthamas, Siripornadulsil, Surasak, Sukon, Peerapol, Sooksawat, Treerat, Siripornadulsil, Wilailak, 2023. Survivability of freeze- and spray-dried probiotics and their effects on the growth and health performance of broilers. Vet. World 16 (9), 1849–1865.

Carvalho, Ana S., Silva, Joana, Ho, Peter, Teixeira, Paula, Malcata, F. Xavier, Gibbs, Paul, 2004. Relevant factors for the preparation of freeze-dried lactic acid bacteria. Int. Dairy J. 14 (10), 835–847.

Celades-Martínez, Jorge, Diago-Mosquera, Melissa E., Peña, Alvaro, García, José, 2026. Shadow fading prediction at 18 GHz through physics guided learning in vegetative corridors. Sci. Rep.

Chawla, Nitesh V., Bowyer, Kevin W., Hall, Lawrence O., Kegelmeyer, W. Philip, 2002. SMOTE: Synthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357.

Chen, Tianqi, Guestrin, Carlos, 2016. XGBoost: A scalable tree boosting system. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining. KDD '16, Association for Computing Machinery, New York, NY, USA, pp. 785–794.

Christ, Maximilian, Braun, Nils, Neuffer, Julius, Kempa-Liehr, Andreas W., 2018. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). Neurocomputing 307, 72–77.

Cortes, Corinna, Vapnik, Vladimir, 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Davis, Catherine, 2014. Enumeration of probiotic strains: review of culture-dependent and alternative techniques to quantify viable bacteria. J. Microbiol. Methods 103, 9–17.

Doran, Pauline M., 2013. Bioprocess Engineering Principles: Second Edition. Elsevier.

Esposito, Alessia, Valentino, Vincenzo, Tagliamonte, Silvia, Sequino, Giuseppina, Vitaglione, Paola, Ercolini, Danilo, De Filippis, Francesca, 2025. Development of a synbiotic dietary supplement containing potential Next Generation Probiotics for modulation of the gut microbiome and metabolome. Curr. Res. Food Sci. 101289.

Ge, Sitong, Han, Jiarun, Sun, Qiaoyu, Zhou, Qingqing, Ye, Zhongdu, Li, Ping, Gu, Qing, 2024. Research progress on improving the freeze-drying resistance of probiotics: A review. Trends Food Sci. Technol. 147, 104425.

Ghiasi, Fatemeh, Hashemi, Seyed Mohammad Bagher, Abedi, Elahe, 2023. Effective enhancement of food oxidative stability induced by Lactobacillus strains: In vitro activity. Food Control 153, 109912.

Hedberg, Sarah H.M., Devi, Sharmila, Duralliu, Arnold, Williams, Daryl R., 2019. Mechanical behavior and structure of freeze-dried cakes. In: Ward, Kevin R., Matejtschuk, Paul (Eds.), Lyophilization of Pharmaceuticals and Biologicals: New Technologies and Approaches. Springer New York, New York, NY, pp. 327–351.

Jannah, Syerina Raihatul, Rahayu, Endang Sutriswati, Yanti, Rini, Suroto, Dian Anggraini, Wikandari, Rachma, 2022. Study of viability, storage stability, and shelf life of probiotic instant coffee Lactiplantibacillus plantarum Subsp. plantarum Dad-13 in vacuum and nonvacuum packaging at different storage temperatures. Int. J. Food Sci. 2022 (1), 1663772.

Junaid, K.P., Kiran, Tanvi, Gupta, Madhu, Kishore, Kamal, Siwatch, Sujata, 2025. How much missing data is too much to impute for longitudinal health indicators? A preliminary guideline for the choice of the extent of missing proportion to impute with multiple imputation by chained equations. Popul. Health Metr. 23 (1), 2.

Kapoor, Sayash, Narayanan, Arvind, 2023. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4 (9), 100804.

Kasper, Julia Christina, Friess, Wolfgang, 2011. The freezing step in lyophilization: Physico-chemical fundamentals, freezing methods and consequences on process performance and quality attributes of biopharmaceuticals. Eur. J. Pharm. Biopharm. 78 (2), 248–263.

Katagiri, Nobuyuki, Kuwajima, Yuya, Kawahara, Hirotaka, Yamashita, Reina, Iritani, Eiji, 2022. Special features of microbial cake under high pressure conditions in microfiltration. Sep. Purif. Technol. 303, 122234.

Khanashyam, Anandu Chandra, Jagtap, Sandeep, Agrawal, Tarun Kumar, Thorakkattu, Priyamvada, Malav, Om Prakash, Trollman, Hana, Hassoun, Abdo, Ramesh, Bharathi, Manoj, Vishnu, Rathnakumar, Kaavya, et al., 2025. Applications of artificial intelligence in the dairy Industry: From farm to product development. Comput. Electron. Agric. 238, 110879.

Kim, Jong, 2019. Multicollinearity and misleading statistical results. Korean J. Anesthesiol. 72 (6), 558–569.

Kırbaş, İsmail, Tuncer, Azim Doğuş, Şirin, Ceylin, Usta, Hüseyin, 2019. Modeling and developing a smart interface for various drying methods of pomelo fruit (Citrus maxima) peel using machine learning approaches. Comput. Electron. Agric. 165, 104928.

Kohavi, Ron, 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proc. 14th Int. Joint Conf. Artif. Intell. IJCAI, Vol. 2, Montreal, Canada, pp. 1137–1145.

Kraus, Michael A., Bartsch, Helen, 2026. Discovery of fatigue strength models via feature engineering and automated eXplainable machine learning applied to the welded transverse stiffener. Int. J. Fatigue 203, 109324.

Kuhn, Max, Johnson, Kjell, 2019. Feature Engineering and Selection: A Practical Approach for Predictive Models. Chapman and Hall/CRC.

Lemaître, Guillaume, Nogueira, Fernando, Aridas, Christos K., 2017. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. 18 (17), 1–5.

Lepot, Mathieu, Aubin, Jean-Baptiste, Clemens, François H.L.R., 2017. Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. Water 9 (10), 796.

Li, Xiangfei, Wang, Menghan, Ding, Meng, Pang, Xinyi, Sun, Jing, Lu, Yingjian, 2025. Effects of complex cryoprotectant on the freeze-drying survival of Lactobacillus acidophilus FMNS-10 and its protective mechanisms. J. Stored Prod. Res. 112, 102646.

Little, Roderick J.A., Rubin, Donald B., 2019. Statistical Analysis with Missing Data. John Wiley & Sons.

Lundberg, Scott M., Lee, Su-In, 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

MacCallum, Robert C., Zhang, Shaobo, Preacher, Kristopher J., Rucker, Derek D., 2002. On the practice of dichotomization of quantitative variables. Psychol. Methods 7 (1), 19–40.

Mählkvist, Simon, Ejenstam, Jesper, Kyprianidis, Konstantinos, 2023. Cost-sensitive decision support for industrial batch processes. Sensors 23 (23), 9464.

Melcher, Michael, Scharl, Theresa, Spangl, Bernhard, Luchner, Markus, Cserjan, Monika, Bayer, Karl, Leisch, Friedrich, Striedner, Gerald, 2015. The potential of random forest and neural networks for biomass and recombinant protein modeling in Escherichia coli fed-batch fermentations. Biotechnol. J. 10 (11), 1770–1782.

Miecznikowski, Antoni, Wierzchowska, Katarzyna, Choińska, Renata, Fabiszewska, Agata, 2026. Tailoring freeze-drying for starter cultures preservation: A case study with Lactiplantibacillus plantarum. Foods 15 (11), 1928.

Molnar, Christoph, König, Gunnar, Herbinger, Julia, Freiesleben, Timo, Dandl, Susanne, Scholbeck, Christian A., Casalicchio, Giuseppe, Grosse-Wentrup, Moritz, Bischl, Bernd, 2022. General pitfalls of model-agnostic interpretation methods for machine learning models. In: Holzinger, Andreas, Goebel, Randy, Fong, Ruth, Moon, Taesup, Müller, Klaus-Robert, Samek, Wojciech (Eds.), International Workshop on Extending Explainable AI beyond Deep Models and Classifiers. Springer International Publishing, Cham, pp. 39–68.

Mondal, Partha Pratim, Galodha, Abhinav, Verma, Vishal Kumar, Singh, Vijai, Show, Pau Loke, Awasthi, Mukesh Kumar, Lall, Brejesh, Anees, Sanya, Pollmann, Katrin, Jain, Rohan, 2023. Review on machine learning-based bioprocess optimization, monitoring, and control systems. Bioresour. Technol. 370, 128523.

Morgan, Charlote A., Herman, N., White, P.A., Vesey, Graham, 2006. Preservation of micro-organisms by drying; A review. J. Microbiol. Methods 66 (2), 183–193.

Nag, Arup, Das, Shantanu, 2013. Improving ambient temperature stability of probiotics with stress adaptation and fluidized bed drying. J. Funct. Food 5 (1), 170–177.

Noufeu, Tchouli, Li, Yueqin, Toure, Ndeye Fatou, Yao, Hui, Zeng, Xiaoqun, Du, Qiwei, Pan, Daodong, 2025. Overview of glycometabolism of lactic acid bacteria during freeze-drying: Changes, influencing factors, and application strategies. Foods 14 (5), 743.

Nti, Isaac, Nyarko-Boateng, Owusu, Aning, Justice, 2021. Performance of machine learning algorithms with different K values in K-fold cross-validation. Int. J. Inf. Technol. Comput. Sci. 13, 61–71.

Oexner, Rafael R., Schmitt, Robin, Ahn, Hyunchan, Shah, Ravi A., Zoccarato, Anna, Theofilatos, Konstantinos, Shah, Ajay M., 2025. Comprehensive benchmarking of machine learning methods for risk prediction modelling from large-scale survival data: A UK biobank study. arXiv preprint arXiv:2503.08870.

Onel, Melis, Kieslich, Chris A., Guzman, Yannis A., Floudas, Christodoulos A., Pistikopoulos, Efstratios N., 2018. Big data approach to batch process monitoring: Simultaneous fault detection and diagnosis using nonlinear support vector machine-based feature selection. Comput. Chem. Eng. 115, 46–63.

Oyinloye, Timilehin Martins, Yoon, Won Byong, 2020. Effect of freeze-drying on quality and grinding process of food produce: A review. Processes 8 (3), 354.

Pagnini, Cristiano, Corleto, Vito Domenico, Martorelli, Michela, Lanini, Claudio, D'Ambra, Giancarlo, Di Giulio, Emilio, Delle Fave, Gianfranco, 2018. Mucosal adhesion and anti-inflammatory effects of Lactobacillus rhamnosus GG in the human colonic mucosa: A proof-of-concept study. World J. Gastroenterol. 24 (41), 4652.

Pandey, Kshipra, Suryavanshi, Mangesh, Patel, Ritu, 2026. Microbial biotechnology in the era of big data and AI. In: Bioinformatics, AI, and Machine Learning in Microbial Drug Development. Elsevier, pp. 91–112.

Papadimitriou, Konstantinos, Alegría, Ángel, Bron, Peter A., De Angelis, Maria, Gobbetti, Marco, Kleerebezem, Michiel, Lemos, José A., Linares, Daniel M., Ross, Paul, Stanton, Catherine, et al., 2016. Stress physiology of lactic acid bacteria. Microbiol. Mol. Biol. Rev. 80 (3), 837–890.

Paul, Prince, Kabbo, Md Kawsarul Islam, Gupta, Rajarshi Das, Rahman Sobuz, Md Habibur, Baral, Prosanjit, Jameel, Mohammed, Abubakar, Sani Aliyu, 2025. Experimental assessment with data-driven machine learning-based prediction of compressive strength of waste natural fiber-reinforced sustainable concrete. Sci. Rep.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Peterson, Brandon W., Sharma, Prashant K., van der Mei, Henny C., Busscher, Henk J., 2012. Bacterial cell surface damage due to centrifugal compaction. Appl. Environ. Microbiol. 78 (1), 120–125.

Quigley, Eamonn M.M., 2011. Gut microbiota and the role of probiotics in therapy. Curr. Opin. Pharmacol. 11 (6), 593–603.

Rathore, Anurag S., Mishra, Somesh, Nikita, Saxena, Priyanka, Priyanka, 2021. Bioprocess control: Current progress and future perspectives. Life 11 (6), 557.

Reddyvari, Ragini, Ren, Yuying, Kosuri, Praveen, Kanike, Eswari, Thankachan, Anjana, Khatri, Kham Bahadur, Amalaradjou, Mary Anne, 2025. Novel probiotics Lactobacillus paracasei and Lactobacillus rhamnosus inhibit Salmonella colonization in layers and prevent egg-borne transmission. Food Control 111781.

Richardson, Ardarion, Walker, Joshua, 2018. Continuous solids-discharging centrifugation: A solution to the challenges of clarifying high–cell-density mammalian cell cultures. Bioprocess Int. 16 (4), 38–47.

Saha, Khokan Kumar, Dulal, Rabin, Schmidtke, Leigh, Zheng, Lihong, Huang, Xiaodi, Clark, Andrew, 2025. Discrimination of vineyard origin and grape cultivars using rapid mass spectrometry and machine learning. Smart Agric. Technol. 101755.

Shi, Xiupeng, Wong, Yiik Diew, Li, Michael Zhi-Feng, Palanisamy, Chandrasekar, Chai, Chen, 2019. A feature learning approach based on XGBoost for driving assessment and risk prediction. Accid. Anal. Prev. 129, 170–179.

Sokolova, Marina, Lapalme, Guy, 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45 (4), 427–437.

Steegen, Sara, Tuerlinckx, Francis, Gelman, Andrew, Vanpaemel, Wolf, 2016. Increasing transparency through a multiverse analysis. Perspect. Psychol. Sci. 11 (5), 702–712.

Strobl, Carolin, Boulesteix, Anne-Laure, Zeileis, Achim, Hothorn, Torsten, 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8 (1), 25.

Thabane, Lehana, Mbuagbaw, Lawrence, Zhang, Shiyuan, Samaan, Zainab, Marcucci, Maura, Ye, Chenglin, Thabane, Marroon, Giangregorio, Lora, Dennis, Brittany, Kosa, Daisy, Debono, Victoria Borg, Dillenburg, Rejane, Fruci, Vincent, Bawor, Monica, Lee, Juneyoung, Wells, George, Goldsmith, Charles H., 2013. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC Med. Res. Methodol. 13 (1), 92.

Touzet, Antoine, Pfefferlé, François, van der Wel, Peter, Lamprecht, Alf, Pellequer, Yann, 2018. Active freeze drying for production of nanocrystal-based powder: a pilot study. Int. J. Pharm. 536 (1), 222–230.

Tulsyan, Aditya, Garvin, Christopher, Ündey, Cenk, 2018. Advances in industrial biopharmaceutical batch process monitoring: Machine-learning methods for small data problems. Biotechnol. Bioeng. 115 (8), 1915–1924.

van der Linden, Aart, Oosting, Simon J., van de Ven, Gerrie W.J., Zom, Ronald, van Ittersum, Martin K., Gerber, Pierre J., de Boer, Imke J.M., 2021. Yield gap analysis in dairy production systems using the mechanistic model LiGAPS-Dairy. J. Dairy Sci. 104 (5), 5689–5704.

Vandewiele, Gilles, Dehaene, Isabelle, Kovács, György, Sterckx, Lucas, Janssens, Olivier, Ongenae, Femke, De Backere, Femke, De Turck, Filip, Roelens, Kristien, Decruyenaere, Johan, Van Hoecke, Sofie, Demeester, Thomas, 2021. Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. Artif. Intell. Med. 111, 101987.

Yang, Ziyi, Lian, Jingyu, Yang, Yuheng, Li, Jiayi, Guo, Weiling, Lv, Xucong, Ni, Li, Chen, Youting, 2025. Selenium enrichment enhances the alleviating effect of Lactobacillus rhamnosus GG on alcoholic liver injury in mice. Curr. Res. Food Sci. 10, 100964.

Zaman, Ferhana, Allan, Corey M., Ho, Sa V., 2009. Ultra scale-down approaches for clarification of mammalian cell culture broths in disc-stack centrifuges. Biotechnol. Prog. 25 (6), 1709–1716.

Zhou, Youran, Aryal, Sunil, Bouadjenek, Mohamed Reda, 2024. Review for Handling Missing Data with special missing mechanism. arXiv preprint arXiv:2404.04905.