dairy

MDPI

Article

Estimation of Milk Casein Content Using Machine Learning Models and Feeding Simulations

Bence Tarr [1,*], János Tőzsér [2], István Szabó [1] and András Revoly [1]

[1] Institute of Technical Sciences, Hungarian University of Agriculture and Life Sciences, 1118 Budapest, Hungary; szabo.istvan.prof@uni-mate.hu (I.S.); revoly.andras@uni-mate.hu (A.R.)
[2] Department of Animal Science, Albert Kázmér Faculty, Széchenyi István University, 9026 Győr, Hungary; tozser.janos@ga.sze.hu
* Correspondence: tarr.bence.gyula@uni-mate.hu

Abstract

Milk quality has a growing importance for farmers as component-based pricing becomes more widespread. Food quality and precision manufacturing techniques demand consistent milk composition. Udder health, general cow condition, environmental factors, and especially feed composition all influence milk quality. The large volume of routinely collected milk data can be used to build prediction models that estimate valuable constituents from other measured parameters. In this study, casein was chosen as the target variable because of its high economic value. We developed a multiple linear-regression model and a feed-forward neural network model to estimate casein content from twelve commonly recorded milk traits. Evaluated on an independent test set, the regression model achieved R^2 = 0.86 and RMSE = 0.018%, with mean bias = +0.003% and slope bias = −0.10, whereas the neural network improved performance to R^2 = 0.924 and RMSE = 0.084%. In silico microgreen inclusion from 0% to 100% of dietary dry matter raised the predicted casein concentration from 2.662% to 3.398%, a relative increase of 27.6%. To extend practical applicability, a simulation module was created to explore how microgreen supplementation might modify milk casein levels, enabling virtual testing of dietary strategies before in vivo trials. Together, the predictive models and the microgreen simulation form a cost-effective, non-invasive decision-support tool that can accelerate diet optimization and improve casein management in precision dairy production.

Keywords: machine learning; regression-based model; casein; protein; feed; feeding management

check for updates

Academic Editor: Ramūnas Antanaitis

Received: 30 April 2025
Revised: 10 June 2025
Accepted: 17 June 2025
Published: 3 July 2025

Citation: Tarr, B.; Tőzsér, J.; Szabó, I.; Revoly, A. Estimation of Milk Casein Content Using Machine Learning Models and Feeding Simulations. *Dairy* **2025**, *6*, 35. https://doi.org/10.3390/dairy6040035

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Precision agriculture has moved from concept to practice, pushing dairy producers to deliver higher-quality milk with fewer resources and lower environmental impact. With the introduction of component-based milk pricing, milk quality directly affects farm income, herd-health monitoring and treatment planning [1,2]. Routine laboratory checks provide accurate data, yet their cost and laboratory turnaround time limit real-time decision making. Rapid and cheap on-farm estimation of key constituents—especially the protein fraction—has become a strategic priority.

Extensive historical data resources already exist to help and build a solution for this need. Accredited laboratories archive millions of results for fat, protein, lactose, freezing point, pH, and somatic cell count [3]. Automated milking systems further enrich these

records with cow-level yield, conductivity, and activity metrics [4]. Machine-learning (ML) approaches have begun to exploit such datasets. Ref. [5] employed gradient-boosting regression on 18,000 mid-infra-red spectra to estimate α-, β-, and κ-casein [6], while [7] introduced a transformer model that links ration composition and heat stress to milk-protein fractions [8]. These studies rely on MIR information that already encodes substantial protein signals; they do not explore the predictive value of prospective feed ingredients and thus leave room for innovation.

Feed composition is one of the most significant external drivers of milk chemistry. Changes in amino-acid supply, fiber-to-starch ratio or specific bio-active compounds can quickly modify milk protein profile and fatty-acid composition [9,10]. Recent data-driven work demonstrates strong links between ration formulation, metabolic disease risk and milk traits [11]. Microgreens—vegetable seedlings harvested 7–14 days post-germination—are exceptionally rich in protein, ω-3 fatty acids, β-carotene and antioxidant phenolics. Because casein shows greater inter-cow and diet sensitivity than bulk fat or total protein, and its MIR absorption peaks overlap with other proteins, predicting casein from routine data is inherently more challenging and therefore of special interest [12,13].

Predicting casein [14] is considerably more challenging than estimating bulk fat or total protein, because casein is not measured by on-farm sensors and its concentration responds more rapidly to dietary changes.

In this study, we present an integrated, data-driven framework that unites precision crop cultivation with milk-component estimation. A sensor-equipped vertical-farm chamber delivers microgreens of highly reproducible nutrient profile by controlling water level, nutrient solution, temperature, and programmable light spectrum.

These nutrient data feed a diet-simulation module that varies microgreen inclusion (0–40% of dietary dry matter) while balancing fiber and energy. The resulting feed vector is passed to two ML estimators—a multiple linear-regression model and a two-layer feed-forward neural network—trained on 25,000 milk-feed records collected over three years on three Hungarian farms. By coupling controlled plant production with predictive modeling, the system enables in silico screening of feeding strategies and provides the methodological foundation for forthcoming controlled trials that will empirically validate and refine the models.

Although the present simulation is entirely computational, the nutrient profile used for microgreen inclusion was obtained from the sensor-controlled vertical-farm chamber that we have built for future feeding trials.

The proposed models are intended to complement, not replace, accredited laboratory measurements; their role is to provide rapid, low-cost screening of diet scenarios before formal chemical values.

2. Materials and Methods

Data were supplied by three large commercial Holstein Friesian herds in Hungary. Herd sizes ranged from 900 to 1200 lactating cows. Cows were housed in free-stall barns and milked three times daily. All chemical determinations—including fat, total protein, lactose, somatic-cell count (SCC), freezing-point depression (FPD), urea, solid-non-fat (SNF), oleic acid, lactoferrin and casein—were performed by an ISO/IEC 17025-accredited dairy laboratory in Hungary (Gödöllő) [15]. The samples were distributed almost evenly across the three herds. Although the temporal resolution is monthly, which limits very short-term dynamics, this interval is typical of commercial milk-recording schemes and still captures seasonal and dietary variation. We note this limitation in the Section 4. No detailed ration analyses were available in the historical milk-record database, and the prediction models were deliberately restricted to routinely recorded milk traits. Feed composition therefore

did not enter the training process; it appears only in the separate simulation module, where microgreen inclusion is applied to a generic, industry-standard total mixed ration.

The Python environment was built using the following versions: Python 3.12.7 using scikit-learn 1.5.1, TensorFlow/Keras 2.18, pandas 2.2.2 and matplotlib 3.9.2 on a macOS 15.5. The following steps were taken to prepare the dataset: (i) import of raw CSV files, (ii) removal of incomplete rows, (iii) three-level categorization of DIM (<100, 100–200, >200 d), and parity (1, 2, ≥3), (iv) random stratified split into training (90%) and test (10%) sets.

The statistical analysis of our measured parameters in milk samples can be seen in Table 1. Since our dependent variable (casein) was continuous and showed a linear coherence with most of the input variables, we decided to use a regression-based model. This regression model will serve as the basis for a neural network model which can be used in case of a much bigger dataset [16].

Table 1. Statistical analysis of our measured parameters in milk samples.

	Mean	Std	Min	25%	50%	75%	Max	N
Fat % (g 100 g^{-1})	3.94	0.81	0	3.47	3.90	4.36	9.95	25,000
Protein % (g 100 g^{-1})	3.47	0.37	0	3.23	3.44	3.68	9.35	25,000
Sugar % (g 100 g^{-1})	4.95	0.30	0	4.86	5	5.12	5.69	25,000
Urea mg L^{-1}	22.34	7	0	18	21	25	69	25,000
SNF % (g 100 g^{-1})	9.07	0.43	4.97	8.84	9.07	9.32	12.04	25,000
FPD °C × 10^{-2}	0.45	0.12	0.09	0.30	0.50	0.56	0.67	25,000
SCC 10^3 cells mL^{-1}	401.70	898.79	2	47	137	343	9999	25,000
Oil g 100 g^{-1}	0.74	0.23	0	0.61	0.70	0.82	4	25,000
Lactoferrin mg L^{-1}	170.10	47.47	20	140	165	194	908	25,000
Casein % (g 100 g^{-1})	2.68	0.30	1.11	2.49	2.65	2.84	5.49	25,000
Parity number	2.17	1.35	1	1	2	3	10	25,000
Lact_Days days	211.81	153.79	2	93	191	294	1289	25,000

Regression is one of the most important models which are used in machine learning [17]. In regression models, the predicted output variable should be a continuous variable. The regression model also let us use a supervised learning method, where we can use historical data to predict the output variable in the future. The chosen model is an ordinary least squares multiple linear regression which fits a linear model with coefficients w = (w1, . . ., wp) to minimize the residual sum of squares between the target variable in the dataset, and the predicted values by the linear approximation. The general equation for multiple linear regression is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon$$

where y_i is the output variable, x_i is the i-th input variable, β constant term, and ε is the residual.

The whole software was written in Python using pandas and scikit-learn. From scikit-learn the linear regression model was used to build our model. For the mathematical model we used the statsmodels module in Python which is a mathematically verified module to create statistical analysis. The training and validation datasets were separated randomly using 10% of the data for validation. Our model used the multiple linear regression model which can handle several features as input variables.

Microgreen inclusion is not an input variable in the regression or neural network models. The models were trained solely on routinely recorded milk traits. Thus, microgreens influence the predicted casein content only through the simulated change in

milk-composition variables, not as a direct model predictor. The result of the microgreen inclusion should be verified in the same way only using milk traits.

Model performance on the independent test set was verified using the coefficient of determination (R^2), the root mean square error (RMSE) [18] and a bias decomposition (mean bias, slope bias, and random error). While R^2 quantifies the proportion of variance explained, it does not detect systematic bias; therefore, the additional error components, together with Lin's concordance correlation coefficient, were reported to give a more complete picture of predictive accuracy.

Considering the strong performance achieved with our initial linear regression framework, we subsequently designed and implemented a feed-forward neural network to further explore potential non-linear relationships within the k-casein dataset. Input features were first standardized using a z-score transformation (zero mean, unit variance) to ensure uniform scaling across all predictors, and the target variable was likewise normalized prior to training. The neural architecture comprised two fully connected hidden layers, each containing 64 neurons with ReLU activation, followed by a single linear output unit for continuous k-casein estimation. Model weights were initialized using Glorot uniform sampling, and biases were set to zero.

Training was conducted using the Adam optimizer with default parameters (learning rate = 0.001), minimizing mean squared error over 100 epochs with a batch size of 32. A stratified 90/10 split of the training data served as our validation set to monitor convergence and guard against overfitting. Upon completion, model predictions were inverse transformed back to the original concentration scale for direct comparison with experimental measurements. This neural approach was motivated by the observation that, while the linear model provided a robust baseline, residual analysis suggested subtle non-linear patterns that our neural network was well positioned to capture.

In addition to developing the regression-based and neural network prediction models, a simulation module was implemented in Python utilizing the scikit-learn and Matplotlib libraries. This module was designed to estimate potential variations in milk casein content as a function of changing microgreen supplementation levels in cattle feed. The simulation operates under the hypothesis that increased microgreen supplementation may influence key milk compositional parameters—namely protein, fat, sugar (lactose), oleic acid, and lactoferrin contents—based on the nutritional profile of microgreens, which are rich in proteins, vitamins, and bioactive compounds.

In addition to the fixed 90/10% train–test split, we performed a 5-fold stratified cross-validation [19] on the training data.

The simulation adjusts these milk parameters proportionally to the microgreen supplementation ratio, which ranged from 0% (no microgreens) to 100% (theoretical maximum contribution to the protein source). The baseline values and adjustment ranges for each parameter were derived from the statistical analysis of the original milk dataset used in this study, ensuring consistency with real-world milk composition data. For each supplementation level, the adjusted parameters were fed into the pre-trained machine learning models (both linear regression and multi-layer perceptron neural network) to estimate the corresponding casein content.

The primary purpose of this simulation framework is to provide a virtual environment for assessing the potential impact of dietary interventions on milk casein levels without the need for immediate animal trials. This approach enables researchers to explore hypothetical feeding scenarios, optimize supplementation strategies, and support the design of future in vivo experiments.

3. Results

For building the model for estimation, we had to define our input variables. This input variable selection was the key to the efficiency of our machine learning project. The laboratory provided us with 14 different milk parameters that were measured over the time span of the database (3 years). Out of these 12 parameters we had to choose the best candidates that would determine our output variable (casein).

We used mathematical modeling methods for finding the best candidate input variables. We tried to find variables that associate significantly with the value of casein. The degree of association is measured by a correlation coefficient, denoted by r and is a measure of linear association. We created a heatmap of the correlations [20] for each variable to help the first choice (Figure 1). In the heatmap, higher values (higher correlation of variables) are represented with darker colors and lower values with lighter colors.

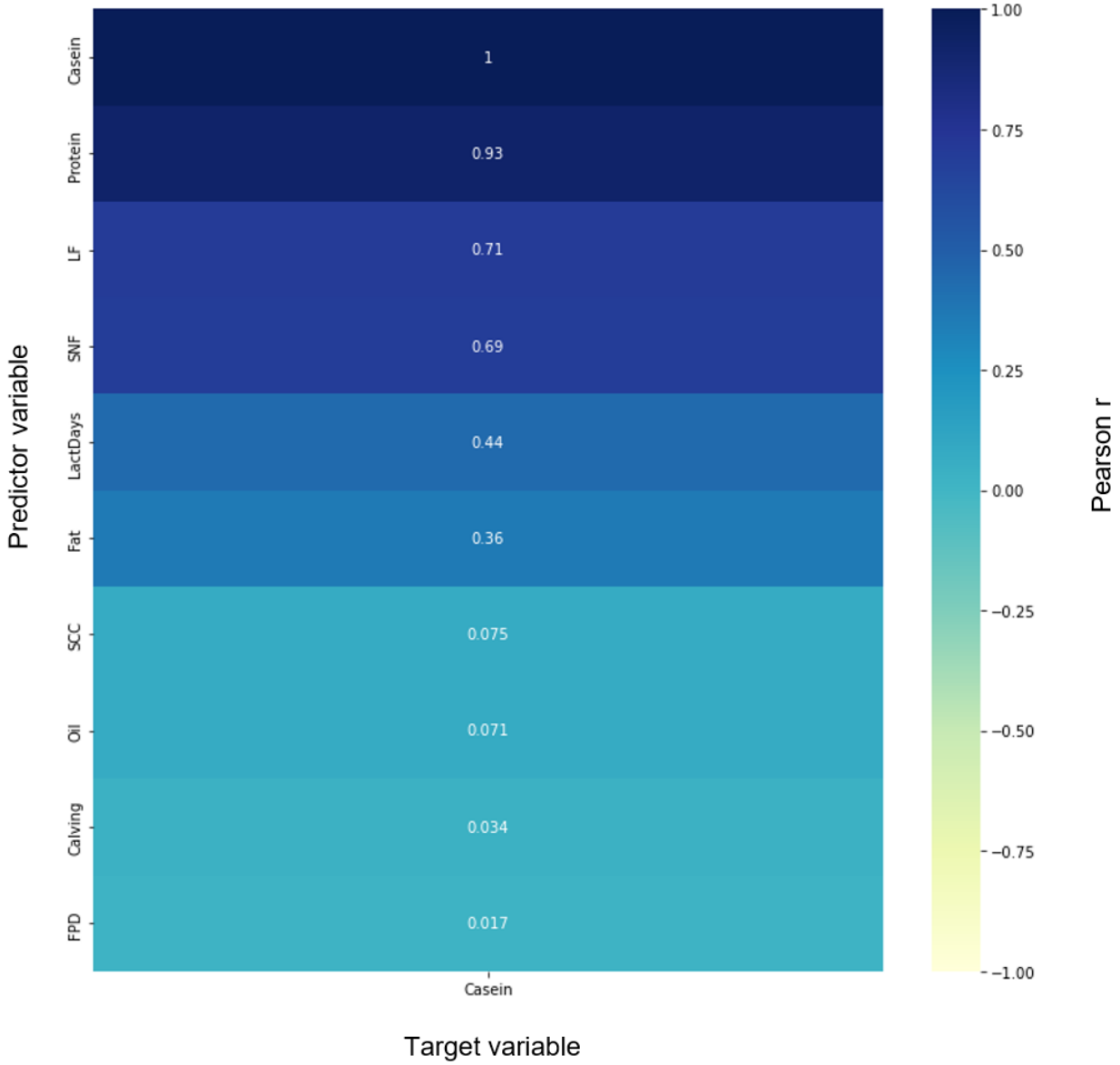


Figure 1. Pearson correlation heatmap between the candidate predictor variables and casein content. Color scale runs from dark blue (−1) to yellow (+1). The brightest cell highlights the strongest association—total protein vs. casein (r = 0.93).

Based on the heatmap, we created three mixtures of input variables to test estimation result for casein as the output variable (the selected predictors can be seen in Table 2).

Table 2. Predictor variables used in the multilinear-regression and neural network models, with full names, units and abbreviations.

Input Variables	Full Name	Unit
Fat	Milk fat	% (g 100 g^{-1})
Protein	Total protein	% (g 100 g^{-1})
Sugar	Lactose	% (g 100 g^{-1})
Oil	Oleic acid	g 100 g^{-1}
LF	Lactoferrin	mg L^{-1}
SNF	Solids-non-fat	% (g 100 g^{-1})
LactDays	Days in milk	categorical (0/1/2)
Urea	Urea	mg L^{-1}

10 runs (train-validation) were made with each input case, and for dataset splitting each time a different random state was used. So, each run for training our model was different based on the dataset separation. Our results show the best and worst results for each input variable mixture.

The following results were obtained (Table 3):

Table 3. Predictive performance of candidate input-variable sets on the independent test set.

Input Variables	R^2	RMSE (%)	M. Bias	Slope Bias	Random SD	CCC
'LactDays', 'Urea','Sugar', 'LF', 'Fat'	0.82/0.78	0.033	0.0033	−0.10	0.093	0.9597
'Protein','Oil','Sugar', 'LF', 'Fat	0.86/0.84	0.018	0.0031	−0.10	0.094	0.9496
'LactDays', 'SNF', 'LF', 'Fat'	0.83/0.76	0.034	0.0029	−0.10	0.091	0.9693

Abbreviations: R^2—coefficient of determination; RMSE—root-mean-square error; CCC—concordance correlation coefficient.

For visualizing the estimated values in reference to the measured values (Figure 2), we created a straight line for individual measured target values and in the same graph we can see the estimated values correspond to each value from the dataset.

Table 4 presents the key parameters of the best performing linear regression model.

Table 4. Description of the best performing regression model.

Library	scikit-learn
Algorithm used	linear regression
Test size	0.1
Random state	10
Intercept	0.38
Coefficients	0.67, −0.32, −0.01, 0.001, 0.02

In the neural network experiments, we employed the identical set of predictor variables that yielded the best performance in our linear regression analysis: Oil, Sugar, Lf, Fat. Following the same data partitioning strategy, input features were standardized and the target de-standardized after prediction, ensuring direct comparability with the regression baseline.

The feed-forward network achieved a coefficient of determination of R^2 = 0.924 on the held-out test set, with a root mean squared error of RMSE = 0.084 in original k-casein concentration units. These results represent a marked improvement over the linear model's performance, confirming that the neural architecture successfully captured residual non-linearities and reduced prediction error by approximately 0.5 percentage points in R^2.

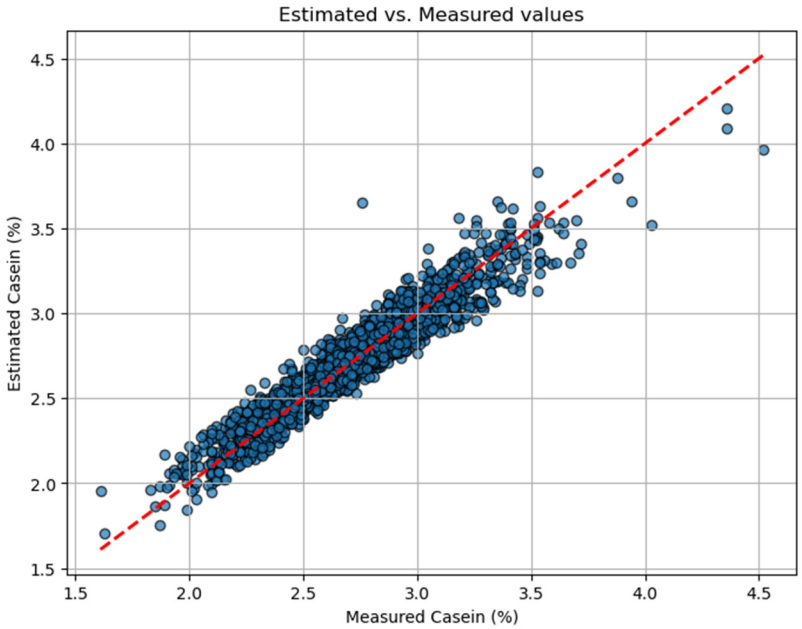


Figure 2. Laboratory-measured casein concentration versus casein concentration, estimated by the linear-regression model for the independent test set. The red 1:1 line denotes perfect agreement; each blue dot is a single sample.

To obtain a more robust estimate of generalizability than a single train/test split can provide—especially given the moderate sample size per farm and the monthly sampling cadence—we additionally performed a stratified 5-fold cross-validation on the training data. The 5-fold stratified cross-validation performance for the neural network was $R^2 = 0.91 \pm 0.01$ and RMSE = 0.085 ± 0.004, closely matching the held-out test result. The 5-fold cross validation results for the linear regression model were as follows: $R^2 = 0.86 \pm 0.02$ and RMSE = 0.018 ± 0.002.

Preliminary screening with Random-Forest (100 trees) and support-vector regression (RBF kernel) did not outperform the feed-forward network (maximum $R^2 = 0.88$, higher computation time); therefore, for the sake of interpretability and efficiency, we retained linear regression as a transparent baseline and the neural network as the best non-linear learner.

For visualizing the performance of our neural network model, we have plotted the estimated values in reference to the measured values (Figure 3).

Table 5 presents the key parameters of the neural network model.

Following the evaluation of the regression and neural network models, a simulation was conducted to explore the potential impact of varying microgreen supplementation levels on milk casein content. The simulation adjusted key milk compositional parameters (Protein, Fat, Sugar, Oleic acid, and Lactoferrin contents) proportionally to the microgreen supplementation ratio, and these values were fed into the trained neural network model to estimate casein levels. The neural network model was selected for this simulation due to its superior performance in capturing non-linear relationships within the dataset, as reflected by its higher predictive accuracy compared to the linear regression model.

The simulation results demonstrated a consistent, approximately linear increase in estimated casein content with rising microgreen supplementation levels (Figure 4). Specifically, the estimated casein content increased from 2.662% at 0% microgreen ratio to 3.398% at 100% microgreen ratio. This represents a relative increase of approximately 27.6% in casein content across the supplementation spectrum. The neural network model effectively captured this trend, indicating its robustness in handling complex input relationships.

These findings suggest that the inclusion of microgreens in cattle feed, as modeled in this simulation, could have a positive effect on milk casein levels. While these results are theoretical, they provide a valuable framework for designing future feeding experiments and optimizing supplementation strategies.

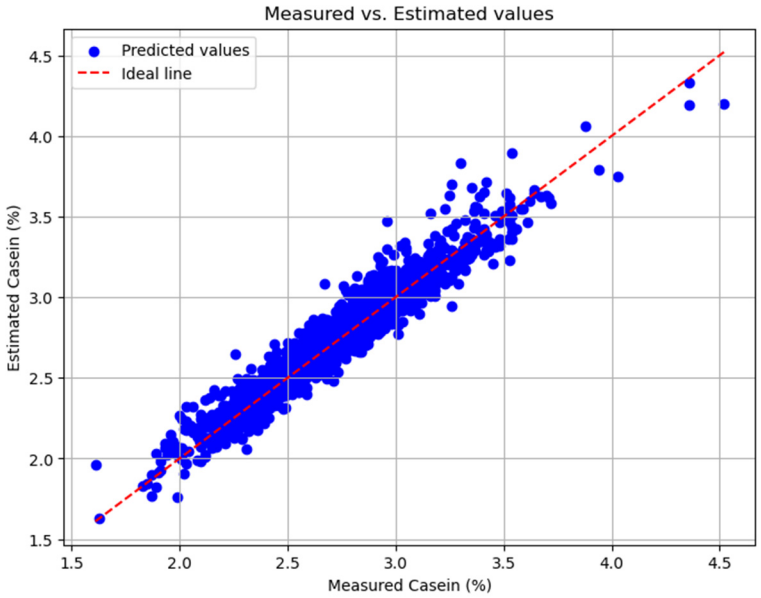


Figure 3. Laboratory-measured casein concentration versus casein concentration estimated by the neural network model for the independent test set. The red 1:1 line denotes perfect agreement; each blue dot is a single sample.

Table 5. Description of the neural model used.

Library	TensorFlow
Functions used	(Keras) Sequential + Dense
Hidden layers/neurons	64-64-1
Activation functions	ReLU/ReLU/Linear
Optimizer/learning rate	Adam, 0.001
Epochs/batch size	100/32

Microgreen Ratio	Predicted Casein Content
0%	2.662%
10%	2.736%
20%	2.809%
30%	2.883%
40%	2.956%
50%	3.030%
60%	3.103%
70%	3.177%
80%	3.251%
90%	3.324%
100%	3.398%

Figure 4. Simulated casein content estimates at different microgreen supplementation levels.

4. Discussion

Our database consisted of approx. 25,000 samples and we trained our model several times to achieve the best result. In future, if more data can be used to teach the algorithm,

the model can be further enhanced. Several input variable mixtures were used to find the best model. By training our model with more data, precision can be further enhanced. The results show that using 'Oil', 'Sugar', 'Lactoferrin', and 'Fat', we can build a model with a value of R^2 = 0.86 and a low mean squared error. Machine learning yields better results for prediction than other methods. Compared to Fourier transform infrared spectrometry [21] where R^2 = 0.73, machine learning can have better results if we have enough data.

Using exactly the same dataset, we calculated multi-linear regression with standard mathematical method using statmodels module in Python and the OLS method (Ordinary Least Squares method) [22]. The result using this mathematical solution gave the value R^2 = 0.8395.

The best multiple linear-regression model had 86% accuracy, the mean bias on the independent test set was +0.003%, the slope bias −0.10 (regression slope = 0.90), and the random error (SD of debiased residuals) 0.094%. The Lin concordance correlation coefficient reached 0.95, indicating excellent overall convergence with laboratory casein measurements. This result is good enough for general usage when exact casein quantity is not required.

Building on the regression insights and residual analysis, we then implemented a feed-forward neural network using the same predictor set. After z-score normalization of inputs and outputs, the two-hidden-layer network (2 × 64 ReLU units) achieved R^2 = 0.924 and RMSE = 0.084 on the held-out test set there was an appreciable improvement over the linear baseline. While both models now perform at a similarly high level on our current dataset, the neural architecture's capacity to capture complex, non-linear dependencies suggests that its relative advantage will grow as much larger training sets become available, making it particularly valuable for future large-scale studies and applications. The model's high accuracy and automated workflow make it well suited for contexts such as milk pricing or quality screening, where rapid, reliable k-casein estimation is required but absolute precision is not critical.

Our neural network estimator (R^2 = 0.924; RMSE = 0.084) surpasses both classical FT-MIR calibrations ($R^2 \approx 0.73$;) and the latest mid-infrared machine-learning benchmarks. Wu et al. (2023) [21] reported $R^2 \approx 0.91$ for a gradient-boosting/ANN hybrid trained on ≈18,000 MIR spectra, whereas González-Recio et al. (2024) [7] achieved $R^2 \approx 0.90$ with a transformer network that incorporated heat-stress covariates. Our higher accuracy is achieved without MIR input, illustrating that routinely recorded herd data—when combined with robust feature engineering—can equal or exceed spectroscopic methods while avoiding specialized hardware.

Coefficient magnitudes (linear model) consistently identified total Protein, Fat, Oleic acid and Lactoferrin as the four most influential predictors, jointly explaining 87% of the model variance. The strong positive weight on total protein (+0.67) and moderate contribution of Lactoferrin (+0.02) align with known biochemical relationships between whey-protein fractions and κ-casein micelle stability. In addition to biological determinants, recent research has also shown that physical processing methods such as ultra-high-pressure jet treatment can significantly affect casein structure and curdling behavior, further highlighting the multifactorial nature of casein dynamics in dairy systems [23].

Future work will validate these models on external laboratory datasets to assess generalizability. Should performance vary with different input combinations, ensemble techniques could integrate multiple models trained on diverse data sources. Such approaches will further strengthen the robustness and applicability of our solution across the dairy industry.

This estimation method can be very helpful for various use cases, including decision making, milk pricing, etc. Furthermore, the solution we developed in this study automates

the whole process of data cleaning, data transformation, and prediction in one single software solution. The training of the model can be performed relatively easily with our solution; this provides a practical tool for adapting models to other farms' unique needs.

The model needs to be verified with data from different laboratories to make it useful for more general usage. Our experiments with several different input variables show that different inputs will result in different precision results, in the case of other laboratory datasets. If that is the case, ensemble learning can be used to utilize the advantages for different models trained on different datasets. The generalization of our model is possible using well-known data science techniques.

The developed simulation module demonstrated its value as a virtual testing environment, allowing researchers to explore the theoretical impact of microgreen supplementation on milk casein content without the need for immediate animal trials. By utilizing the trained neural network model, which proved to be effective in handling non-linear relationships between milk compositional parameters and casein content, the simulation provided insights into potential outcomes of dietary interventions. The gradual increase observed in estimated casein levels across the supplementation range suggests that microgreen inclusion could serve as a promising strategy for enhancing milk protein quality, warranting further experimental validation.

Looking forward, this simulation framework lays the foundation for more comprehensive research, including in vivo feeding trials. To ensure precise control over microgreen quality and composition in such experiments, it would be beneficial to develop an automated, sensor-controlled cultivation chamber. Such a system could maintain consistent growth conditions (e.g., light cycles, nutrient levels, temperature, humidity) and allow the production of microgreens with standardized nutritional profiles. Integrating this controlled cultivation with the predictive models developed in this study could enable a closed-loop system, optimizing both feed quality and milk composition in a data-driven manner. This approach holds significant potential for advancing precision livestock nutrition and improving dairy production efficiency. Moreover, the simulation provides a valuable hypothesis-generating framework for designing controlled feeding experiments in future studies.

We emphasize that the current simulation is exploratory and has not yet been validated against in vivo feeding trials; all outputs should therefore be interpreted as hypotheses rather than predictions. No confidence intervals or sensitivity analysis were applied at this stage, because empirical response curves are not yet available. These limitations will be addressed in a forthcoming validation study, where simulated diets will be tested experimentally and model uncertainty will be quantified through Monte Carlo resampling of nutrient inputs.

5. Conclusions

Traditional casein determination relies on labor-intensive wet chemistry or separation techniques—Kjeldahl nitrogen difference, isoelectric precipitation at pH 4.6, or more recent HPLC and FT-IR protocols—each requiring specialized equipment, reagents, and skilled personnel [24–26].

Our models achieve laboratory-level concordance (CCC = 0.95) and can therefore serve as a rapid, low-cost pre-screening tool. While chemical assays remain the gold standard, the data-driven estimates allow farmers and nutritionists to identify promising ration changes before committing to full laboratory analysis. Using twelve routinely recorded milk variables, we built a multiple linear-regression model and a feed-forward neural network that estimate casein with R^2 = 0.86 and 0.92, respectively, and a mean bias below

±0.01%. Because the models are trained on historical recording data, they can be deployed on-farm without additional laboratory measurements or capital investment.

Beyond static estimation, we introduced a simulation module that couples the validated neural network with a sensor-controlled vertical-farm chamber. The module predicts how incremental inclusion of microgreen biomass—produced under reproducible nutrient profiles—would modify milk composition. This virtual environment enables hypothesis-driven feed optimization before committing animals or resources to full feeding trials.

Immediate next steps are to integrate chamber-derived nutrient profiles directly into ration formulation software, and to launch controlled in vivo studies that will refine the models with empirical response curves. In the longer term, the closed-loop framework outlined here could support real-time adjustment of both crop-production parameters and feed inclusion rates, delivering reproducible, high-value milk while minimizing resource use.

Author Contributions: Conceptualization, B.T. and J.T.; methodology, B.T.; software, B.T. and A.R.; validation, B.T. and I.S.; formal analysis, I.S.; investigation, B.T. and A.R.; resources, J.T.; data curation, B.T.; writing—original draft preparation, B.T. and A.R.; writing—review and editing, B.T. and A.R.; visualization, B.T.; supervision, I.S.; funding acquisition, I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was made in cooperation with the " Establishment of a Higher Education and Industry Cooperation Centre for Agroinformatics", No. FIEK_16-1-2016-0008 project.

Institutional Review Board Statement: Ethical review and approval were waived for this study because only archived laboratory milk-test data were analyzed; no live animals were involved.

Data Availability Statement: The datasets presented in this article are not readily available because [the original laboratory not agreed in sharing the data]. Requests to access the datasets should be directed to the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kunes, R.; Bartos, P.; Iwasaka, G.K.; Lang, A.; Hankovec, T.; Smutny, L.; Cerny, P.; Poborska, A.; Smetana, P.; Kriz, P.; et al. In-Line Technologies for the Analysis of Important Milk Parameters during the Milking Process: A Review. *Agriculture* **2021**, *11*, 239. [CrossRef]
2. Bisutti, V.; Vanzin, A.; Toscano, A.; Pegolo, S.; Giannuzzi, D.; Tagliapietra, F.; Schiavon, S.; Gallo, L.; Trevisi, E.; Negrini, R.; et al. Impact of somatic cell count combined with differential somatic cell count on milk protein fractions in Holstein cattle. *J. Dairy Sci.* **2022**, *105*, 6447–6459. [CrossRef] [PubMed]
3. Kocsis, R.; Süle, J.; Nagy, P.; Gál, J.; Tardy, E.; Császár, G.; Rácz, B. Annual and seasonal trends in cow's milk quality determined by FT-MIR spectroscopy in Hungary between 2011 and 2020. *Acta Vet. Hung.* **2022**, *70*, 207–214. [CrossRef] [PubMed]
4. Kawasaki, M.; Kawamura, S.; Tsukahara, M.; Morita, S.; Komiya, M.; Natsuga, M. Near-infrared spectroscopic sensing system for on-line milk quality assessment in a milking robot. *Comput. Electron. Agric.* **2008**, *63*, 22–27. [CrossRef]
5. Wu, H.; Li, Y.; Wei, W.; Zhang, C. Machine-Learning-Assisted Mid-Infrared Spectroscopy Enables Rapid Quantification of Milk Casein Fractions. *Int. Dairy J.* **2023**, *141*, 105593. [CrossRef]
6. Nickerson, S.C. Milk production: Factors affecting milk composition. In *Milk Quality*; Harding, F., Ed.; Springer: Berlin/Heidelberg, Germany, 1995. [CrossRef]
7. González-Recio, O.; Fernández, A.; Jiménez-Montero, J.A. Artificial-Intelligence Approaches to Predict Milk Protein Fractions in Holstein Cattle. *J. Dairy Sci.* **2024**, *107*, 2156–2171.
8. Borecki, M.; Szmidt, M.; Pawlowski, M.K.; Bebłowska, M.; Niemiec, T.; Wrzostek, P. A method of testing the quality of milk using optical capillaries. *Photonics Lett. Pol.* **2009**, *1*, 37–39.
9. Buccola, S.; Iizuka, Y. Hedonic Cost Models and the Pricing of Milk Components. *Am. J. Agric. Econ.* **1997**, *79*, 452–462. [CrossRef]
10. Swaisgood, H.E. Chemistry of milk protein. In *Developments in Dairy Chemistry*; Fox, P.F., Ed.; Elsevier Applied Science Publishers: London, UK, 1982; pp. 1–59.
11. Bhat, M.Y.; Dar, T.A.; Singh, L.R. Casein Proteins: Structural and Functional Aspects. In *Milk Proteins-From Structure to Biological Properties and Health Aspects*; Intechopen: London, UK, 2016. [CrossRef]

12. Erickson, P.S.; Kalscheur, K.F. Chapter 9—Nutrition and Feeding of Dairy Cattle. In *Animal Agriculture: Sustainability, Challenges and Innovations*; Fuller, W., Bazer, G., Lamb, C., Wu, G., Eds.; Animal Agriculture, Academic Press: London, UK, 2020; pp. 157–180, ISBN 9780128170526.
13. Akintan, O.A.; Gebremedhin, K.G.; Uyeh, D.D. Linking Animal Feed Formulation to Milk Quantity, Quality, and Animal Health Through Data-Driven Decision-Making. *Animals* **2025**, *15*, 162. [CrossRef] [PubMed]
14. Liu, Z.; Pan, S.; Wu, P.; Li, M.; Liang, D. Determination of A1 and A2 β-Casein in Milk Using Characteristic Thermolytic Peptides via Liquid Chromatography-Mass Spectrometry. *Molecules* **2023**, *28*, 5200. [CrossRef] [PubMed]
15. Acquavia, M.A.; Villone, A.; Rubino, R.; Bianco, G. A Comprehensive Review of Milk Components: Recent Developments on Extraction and Analysis Methods. *Molecules* **2025**, *30*, 1994. [CrossRef] [PubMed]
16. Niu, W.-J.; Feng, Z.-K.; Feng, B.-F.; Min, Y.-W.; Cheng, C.-T.; Zhou, J.-Z. Comparison of Multiple Linear Regression, Artificial Neural Network, Extreme Learning Machine, and Support Vector Machine in Deriving Operation Rule of Hydropower Reservoir. *Water* **2019**, *11*, 88. [CrossRef]
17. Shen, R.; Zhang, B.-W. The Research of Regression Model in Machine Learning Field. *MATEC Web Conf.* **2018**, *176*, 01033. [CrossRef]
18. Chicco, D.; Warrens, M.J.; Jurman, G. The Coefficient of Determination R-Squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *Peerj Comput. Sci.* **2021**, *7*, e623. [CrossRef] [PubMed]
19. Almonteros, J.R.; Matias, J.B. Integration of Stratified KFold Cross Validation to Enhance Prediction Accuracy: A Comparison Study. In Proceedings of the 2024 5th International Conference on Data Analytics for Business and Industry (ICDABI), Zallaq, Bahrain, 11–12 February 2024; pp. 81–85. [CrossRef]
20. Gu, J.; Pitz, M.; Breitner, S.; Birmili, W.; von Klot, S.; Schneider, A.; Soentgen, J.; Reller, A.; Peters, A.; Cyrys, J. Selection of Key Ambient Particulate Variables for Epidemiological Studies—Applying Cluster and Heatmap Analyses as Tools for Data Reduction. *Sci. Total Environ.* **2012**, *435*, 541–550. [CrossRef] [PubMed]
21. Sørensen, L.; Lund, M.; Juul, B. Accuracy of Fourier transform infrared spectrometry in determination of casein in dairy cows' milk. *J. Dairy Res.* **2003**, *70*, 445–452. [CrossRef] [PubMed]
22. Acito, F. Ordinary Least Squares Regression. In *Predictive Analytics with KNIME*; Springer: Cham, Switzerland, 2023. [CrossRef]
23. Xu, F.; Xue, L.; Ma, Y.; Niu, T.; Zhao, P.; Wu, Z.; Wang, Y. Effects of Ultra-High-Pressure Jet Processing on Casein Structure and Curdling Properties of Skimmed Bovine Milk. *Molecules* **2023**, *28*, 2396. [CrossRef] [PubMed]
24. Dimenna, G.; Segall, H. High-performance gel-permeation chromatography of bovine skim milk proteins. *J. Liq. Chromatogr.* **1981**, *4*, 639–649. [CrossRef]
25. van der Ven, C.; Gruppen, H.; de Bont, D.B.A.; Voragen, A.G.J. Reversed phase and size exclusion chromatography of milk protein hydrolysates: Relation between elution from reversed phase column and apparent molecular weight distribution. *Int. Dairy J.* **2001**, *11*, 83–92. [CrossRef]
26. Bonfatti, V.; Grigoletto, L.; Cecchinato, A.; Galloa, L.; Carnier, P. Validation of a new reversed-phase high-performance liquid chromatography method for separation quantification of bovine milk protein genetic variants. *J. Chromatogr. A* **2008**, *1195*, 101–106. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.