revista interfaces
Artigos Originais
V. 13, N. 3 (2025) | ISSN 2317-434X

IS IT POSSIBLE TO DETECT WATER ADULTERATION IN RAW MILK THROUGH HIDDEN PATTERNS IN ITS CENTESIMAL COMPOSITION?

DOI: 10.16891/2317-434X.v13.e3.a2025.id2110

Recebido em: 25.07.2024 | Aceito em: 11.03.2025

Rafael Fagnani[a]*, Ricardo César Tavares Carvalho[b], Aline Romano Cunha[a], Josiane Ito Eleodoro[a], Elsa Helena Walter de Santana[a]

Universidade Anhanguera Unopar, Londrina – PR, Brasil[a]
Universidade de Cuiabá – UNIC, Cuiabá – MT, Brasil[b]
****E-mail: rafaelfagnani@hotmail.com***

ABSTRACT

Tampering with raw milk by adding water is a prevalent form of food fraud driven by economic motives. Although the freezing point can detect this adulteration, Brazil and many other regions only mandate this analysis for bulk-tank samples, which can mask the fraud when raw milk is pooled. However, individual compositional analysis of milk from producers is conducted monthly. This study aimed to leverage compositional analysis combined with artificial intelligence to predict water addition. Between 2021 and 2023, 278 raw milk samples from a dairy plant were analyzed using mid-infrared spectroscopy to determine fat, protein, lactose, urea, ash, total solids, and non-fat solids. Simultaneously, these samples were assessed for freezing point, excluding samples with a freezing point above -0.555°CH from the analysis. Of the remaining 251 samples, the Naive Bayes classifier was applied to distinguish between samples with a freezing point greater than -0.530°CH (indicating water addition) and those between -0.531°CH and -0.555°CH (normal). The classifier was trained with 75% of the data, and its performance was evaluated for sensitivity and specificity. The Naive Bayes classifier demonstrated 8% sensitivity, 98% specificity, and 38% positive predictive value. Limitations included a small sample size and the low frequency (15%) of adulterated samples. Future studies with larger sample sizes could improve predictions. Despite its low sensitivity, this algorithm can assist dairy processing plants in identifying potential cases of water addition, aiding in the gradual elimination of such adulteration.

Keywords: Naive Bayes; Freezing point; Food fraud**.**

revista interfaces Artigos Originais
V. 13, N. 3 (2025) | ISSN 2317-434X

INTRODUCTION

The quality and identity of refrigerated raw milk is one of the top priorities in the dairy industry. The authenticity of the raw material is crucial both for the productivity of industrialized products and for compliance with current regulations. One of the most common forms of milk adulteration is watering, an economically motivated action that results in reduced industrial yield. It is estimated that the occurrence of fraud by adding water to refrigerated raw milk ranges between 5% and 20% (SANTOS, 2024; GASPAROTTO *et al*., 2020).

Milk watering can be detected through various laboratory techniques, with cryoscopy being one of the most used. Cryoscopy measures the freezing point of milk, which is altered when water is added. Although effective, this method requires specific equipment and laboratory analysis, which can be impractical for continuous and large-scale monitoring. In this context, the need arises for predictive methods that can identify adulteration more practically and efficiently.

In the past five years, the application of big data and machine learning has shown significant potential in various areas of science and industry. These methods allow the analysis of large volumes of data and the identification of complex patterns that may not be visible through traditional analyses (KAVITHA; DEEPA, 2021). Several advanced techniques have been developed, such as the use of machine learning algorithms like Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine, and Gradient Boosting Machine to classify pure and adulterated milk (NETO *et al*., 2019).

Combined with these machine learning methods, Fourier Transform Infrared Spectroscopy (FTIR), a type of non-destructive analysis commonly found in articles using models to distinguish between pure and adulterated milk samples, achieving high classification accuracies of up to 98.76% (MONTOYA; RODRIGUEZ; LI, 2022). Another analysis technology is Wireless Power Transfer (WPT), which has been explored for non-invasive classification of milk adulterated with different volumes of water, achieving average accuracies of up to 97.6% (VIDHYA *et al*., 2023). These approaches demonstrate that analysis technologies associated with data processing can be an ally in combating milk adulteration.

This type of statistical approach is possible because the centesimal composition of milk, which includes parameters such as protein, fat, lactose, urea, and total solids, offers a diverse source of data for analysis. By analyzing these variables in conjunction with the milk's freezing point, it is possible to develop predictive models that identify adulteration patterns. Integrating this information through machine learning algorithms can enable the creation of highly accurate automatic systems for detecting watering.

This technology is already a reality in Brazil. Clínica do Leite, one of the laboratories accredited by the Brazilian Milk Quality Network (RBQL), in partnership with FOSS Analytics, developed and implemented a system to monitor the presence of 10 adulterants in raw milk. The method developed was the calibration and validation of equipment already used in the routine of some industries and milk analysis laboratories, the MilkoScan FT, which uses Fourier transform infrared spectroscopy. Consequently, the analysis of adulterants does not require additional sampling or analyses and can be incorporated into the monthly and mandatory monitoring for somatic cells and bacterial count. The statistical approach involves targeted and untargeted models, which together can identify and subsequently quantify adulterants. The operating system offers 84% sensitivity and 100% specificity for all evaluated adulterants, with the best sensitivity for corn starch (>98%), sodium bicarbonate (100%), sodium citrate (99%), and formaldehyde (>84%) (COITINHO *et al*., 2017).

Thus, fraud detection through machine learning algorithms can be a beneficial tool for controlling the quality of refrigerated raw milk, helping to ensure its authenticity and quality, with advantages over the isolated use of traditional methodologies, such as speed and low operational cost. The aim of this study was to develop and evaluate predictive systems for watering in refrigerated raw milk based on big data from the centesimal composition, which can meet the commercial interests of dairies and consumers' food integrity.

In this sense, we aimed to use the compositional analysis of raw milk for predicting water addition using artificial intelligence.

MATERIAL AND METHODS

revista interfaces
Artigos Originais
V. 13, N. 3 (2025) | ISSN 2317-434X

Construction of Databases

The database was constructed from the analyses of centesimal composition associated with the freezing point of refrigerated raw milk samples. Sampling was sourced from two dairies in the northern region of Paraná, which, to comply with regulations, periodically conducted:

• Monthly collection of refrigerated raw milk samples from individual or communal tanks for centesimal composition analysis at laboratories accredited by the Ministry of Agriculture.

• Daily collection of refrigerated raw milk samples from individual or communal tanks for tracking.

Using the logistics of each industrial establishment, the samples for centesimal composition evaluation (200 mL) were stored in polyethylene bottles with bronopol and sent under refrigeration in insulated boxes with recyclable ice to the laboratory of the Association of Dutch Cattle Producers and Breeders (APCBRH) in Curitiba-PR, where the percentages of urea, fat, protein, lactose, non-fat solids, and total solids were measured using mid-infrared spectroscopy (International IDF standard 141c, 2000).

Simultaneously, the corresponding samples collected for tracking had their freezing point index evaluated according to the official methodology prescribed by the Ministry of Agriculture (BRAZIL, 2017).

Development and Training of a Predictive Watering System (Machine Learning)

The freezing point values were grouped into two categories: (1) greater than or equal to -0.530°H and (2) below -0.531°H. Although values of -0.530°H are still within legal limits, they were included as indicative of watering due to representing a situation of concern.

Considering that the focus of the study is to predict cases of watering, samples with cryoscopy below -0.555°H were excluded from statistical analyses, as they skew machine learning training and impair the sensitivity and specificity of the generated algorithms.

The Pearson Correlation and three machine learning techniques (Logistic Regression, Support Vector Machines (SVM), and Naive Bayes algorithm) were used to predict water adulteration in milk.

In logistic regression, each variable underwent univariate analysis to assess individual contributions, followed by the removal of non-significant factors ($p>0.05$) using stepwise backward elimination. Validation of the logistic regression included Pearson, Deviance, and Hosmer-Lemeshow tests.

The Naïve Bayes and SVM training were done with 75% of the whole data (randomly chosen). To evaluate the classification efficiency, sensitivity and specificity were calculated. Sensitivity was defined as the percentage of samples classified as "watered" with a cryoscopic index ≥ -0.530°C, while specificity was the percentage of samples classified as normal with a cryoscopic index between -0.531°C and -0.555°C. Positive predictive value was the probability of a sample being "watered" with a cryoscopic index ≥ -0.530°C, and negative predictive value was the probability of a sample being normal with a cryoscopic index between -0.531°C and -0.555°C.

The Support Vector Machine (SVM), a binary linear classifier, was used for data mapping on a plane and the subsequent search for a separating line (hyperplane) between "watered" and "normal" samples.

The Naive Bayes algorithm was also applied as a probabilistic classifier, seeking independent patterns between cryoscopy and the centesimal parameters of each sample, allowing predictions of watering results (cryoscopy greater than or equal to -0.530 °H) based on centesimal data.

In both models, two training sets were used: (1) with 75% of randomly chosen samples and; (2) with the first 200 entries, composed of 50% with cryoscopy greater than or equal to -0.530°H ("watered") and 50% between -0.531°H and -0.555°H (normal).

Evaluation of Algorithm Sensitivity and Specificity

The applicability of the algorithms was evaluated through the concepts of sensitivity, specificity, positive predictive value, and negative predictive value. Based on the classification performed by the algorithms, 2x2 tables were constructed, inputting observed cases (which will use cryoscopy as reference) against predicted cases (those classified according to the algorithm).

revista interfaces Artigos Originais
V. 13, N. 3 (2025) | ISSN 2317-434X

Sensitivity was defined as the percentage of samples classified as “watered” among those with a cryoscopic index greater than or equal to -0.530°H. Specificity was defined as the percentage of samples classified as normal among those with a cryoscopic index between -0.531°H and -0.555°H. Positive and negative predictive values were defined as the probability of a sample being “watered” among samples with a cryoscopic index greater than or equal to -0.530°H and of being “normal” among samples with a cryoscopic index between -0.531°H and -0.555°H, respectively.

All analyses were conducted using Statistica 13.0 software.

RESULTS AND DISCUSSION

The Pearson correlation showed association ($p<0.05$) between cryoscopy and percentage of urea (r= -0.17), protein (r= 0.18), and lactose (r= -0.13). Thus, the higher the amounts of urea and lactose, the lower the freezing points of the samples, while higher protein amounts result in higher freezing points. Although significant, the correlation coefficient was low. This allows us to infer that less than 20% of the cryoscopy variation cases are associated with fluctuations in urea, protein, or lactose values. The weak association between the variables in the studied sample did not allow for the prediction of adulteration with the reliability required.

The results of the logistic regression also indicated that the percentage of urea was related to the probability of adulteration (p= 0.02). However, the model was not adequate for practical application since most "watered" and "normal" samples shared the same urea values. Thus, the model achieved no correct predictions (0% sensitivity) in identifying potential cases of adulteration. The model evaluation using the Hosmer-Lemeshow test also showed poor fit, and the Nagelkerke coefficient was extremely low (0.04). In this case, less than 4% of the variations in urea values were due to extrapolation of the freezing point. Therefore, the use of logistic regression for predicting potential adulteration in raw milk was not suitable.

The results obtained using the Support Vector Machine (SVM) were unsatisfactory, with an accuracy of 5%. This value indicates insufficient performance for an adulteration detection method that requires high reliability.

Sensitivity was 6%, meaning the model correctly identified only 6% of the “watered” samples. This implies that 94% of the adulterated samples were not detected, being classified as “normal” and potentially compromising the quality of the analyzed milk.

The model's specificity was 70%, indicating that 30% of the “normal” samples were incorrectly classified as “watered.” Although specificity is slightly better than sensitivity, it is still far from ideal, especially for an application where the detection of false positives should be minimal.

The F1-Score, which is a combined metric of precision and recall, was 62%. This value reflects the overall poor performance of the classifier, showing that the balance between sensitivity and specificity was not adequate.

For the specific detection of “watered” samples, the SVM correctly identified 6% of the samples. However, 56 “watered” samples were incorrectly classified as “normal,” representing a significant failure in the detection system.

In detecting “normal” samples, the classifier performed slightly better but was still unsatisfactory. Of the 139 “normal” examples, 97 were correctly identified, while 42 were incorrectly identified as “watered.”

These results indicate that the use of SVM for detecting watering in refrigerated raw milk was not effective. The rates of false negatives and false positives were high, compromising the practical utility of the model. Significant improvements to the model or the consideration of more robust classification alternatives are needed to achieve reliable results in milk adulteration detection.

The Naive Bayes classifier showed a sensitivity of 8%, specificity of 98%, and a positive predictive value of 38%. These values indicate that while the algorithm had a low detection rate for “watered” samples, it was highly effective in correctly identifying “normal” samples and had moderate accuracy in predicting fraudulent cases.

A sensitivity of 8% shows that Naive Bayes correctly identified only a small fraction of adulterated samples, indicating a significant limitation in fraud detection. On the other hand, the high specificity of 98% suggests that the algorithm rarely classifies normal samples as adulterated, minimizing false positives. The positive predictive value of 38% indicates that, of the samples classified as “watered” by the algorithm, 38% were actually adulterated.

Despite the low sensitivity, this algorithm had a high specificity and thus could help the milk dairy processing plant to identify 4 out of 10 cases predicted as adulterated, bring useful not for screening purposes, but to gradually banish individual cases of true water addition in raw milk

These results, although still insufficient for a complete practical application, indicate the potential of Naive Bayes as an auxiliary tool in quality control. With further improvements, the algorithm could become more effective in detecting adulterated samples, contributing to the integrity and safety of refrigerated raw milk.

In this study, limitations were due to low sampling and low frequency (4% of samples with cryoscopy above -0.530). To minimize this issue, it is recommended to classify samples with cryoscopy equal to -0.530 as alert, as it improves Naive Bayes prediction.

Future studies could achieve better predictions with sample sizes exceeding 1,000 cases. In practice, it is recommended that dairies frequently update the database, as a larger sample size will enhance the robustness of the Bayesian model. It is also advisable for each dairy or milk basin to develop its own model, as there may be a loss of robustness in different contexts.

Our findings can contribute to the scientific community by showcasing the applications and limitations of artificial intelligence applied to raw milk compliance. Several dairy plants can benefit from our findings, as water adulteration occurs in about 10% of individual producer samples. These models can be transformed into software, aiding in the swift and simple identification of non-conformities in the dairy supply chain without incurring additional laboratory analysis costs. Therefore, this research is contemporary, relevant, and addresses the commercial interests of dairies and the food integrity concerns of consumers.

CONCLUSION

In conclusion, the link between milk's centesimal composition and its freezing point is intricate, posing difficulties in creating a dependable algorithm for predicting water adulteration in raw milk. Artificial intelligence methods have outperformed the statistical techniques reviewed, uncovering hidden patterns in the data that typically led to algorithms with high specificity but low sensitivity. Among these, the Naive Bayes classifier performed the best, showing sufficient specificity to serve as a preliminary alert tool. It can help identify cases that may need on-site technical evaluations to verify potential adulteration, with a 40% chance of confirming these suspicions.

REFERENCES

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. **Manual de métodos oficiais para análise de alimentos de origem animal** (2. ed.). Secretaria de Defesa Agropecuária. Ministério da Agricultura, Pecuária e Abastecimento. Brasília: MAPA, 2019.

COITINHO, T. B.; OLIVEIRA, J. S.; SILVA, A. M.; SOUZA, L. B. Adulteration identification in raw milk using Fourier transform infrared spectroscopy. **Journal of Food Science and Technology***,* v. 54, n. 10, p. 2394-2402, 2017.

GASPAROTTO, P. H.; SILVA, A. J.; SANTOS, M. A. Coleta de dados do índice crioscópico de leite cru refrigerado produzido na microrregião de Ji-Paraná–Rondônia. **Revista Veterinária em Foco***,* v. 17, n. 2, 2020.

INTERNATIONAL DAIRY FEDERATION. (2000). *International IDF standard 141C*. International Dairy Federation.

KAVITHA, P. V.; DEEPA, P. V. (2021). A comparative analysis of the machine learning methods for milk adulteration detection. *In*: **AIP Conference Proceedings** (Vol. 2380, No. 1, pp. 1-8). AIP Publishing.

MONTOYA, N. V.; RODRIGUEZ, D.; LI, C. Wireless Power Transfer Sensing Approach for Milk Adulteration Detection Using Supervised Learning. In: **2022 IEEE**

revista interfaces Artigos Originais
V. 13, N. 3 (2025) | ISSN 2317-434X

Radio and Wireless Symposium (RWS), pp. 131-134, 2022. IEEE.

NETO, H. A.; FIGUEIREDO, J. F.; CASTRO, M. P. On the utilization of deep and ensemble learning to detect milk adulteration. **BioData Mining**, v. 12, n. 1, p. 1-13, 2019.

SANTOS, Gilvania dos. **Detecção de fraudes em leite cru refrigerado na região do Alto Sertão Sergipano**. 2024. 45 f. Trabalho de Conclusão de Curso (Graduação em Agroindústria) - Universidade Federal de Sergipe, Nossa Senhora da Glória, 2024.

Vidhya, S.; Kumari, S.; Rani, M. (2023). Milk quality prediction using supervised machine learning technique. In: **International Conference on Emerging Trends in Expert Applications & Security** (pp. 267-278). Springer Nature Singapore.