ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

Predicting Milk Quality with Machine Learning

[1] K. Rajkumar, [2] K. Kirthana,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract

Milk is an essential part of everyone's diet. There shouldn't be any fillers in premium milk. All across society, you may find dairy products for sale. However, the evaporated milk is permanently changed since the local milk dealers utilize a variety of adulterants. Serious health problems might arise from using spoiled milk. According to the final results of the National Milk Safety and Quality Survey (NMSQS), which were announced on October 18th by the Food Safety and Standards body of India (FSSAI), the leading food safety body in the country, the milk that is easily accessible in India is "mostly safe." A 68.4 percent contamination rate in India's milk supply was found in a recent FSSAI study. No machine or method has been found yet that can guarantee milk quality. Pasteurized milk has been treated to kill hazardous bacteria, while unpasteurized milk has not been. Salmonella, Campylobacter, Cryptosporidium, Escherichia coli, Listeria, and Brucella are only few of the harmful organisms that might be present in contaminated raw milk. Your family's health is in grave danger from these microbes. Determining milk quality by manually evaluating its numerous elements is no easy task. Machine learning-powered analysis and discovery may be of great assistance in this attempt. In this article, we build a system to predict milk quality using machine learning. There has been a 99.99% success rate in classifying data using the suggested method.

Machine learning, label encoding, support vector machines, random forests, and milk quality prediction

1.Introduction

For all people, milk is an essential part of their daily routine. Therefore, for optimal health, it is recommended to drink high-quality milk. Milk is an example of a perishable product. Significant financial losses may occur when even a single gram of milk with poor quality or structure ruins tons of milk. Milk that has gone bad may quickly become a breeding ground for millions of germs. [1] Thus, situations may emerge when the use of milk or dairy products poses a threat to human health. Contamination of food causes over 48 million medical cases annually in the United States. Developing countries like India face many challenges at once due to the fact that poorly maintained dairy products may spread multiple diseases and lead to epidemics of brucellosis, listeriosis, TB, etc. A 68.4 percent contamination rate in India's milk supply was found in a recent FSSAI study. Controlling the spread of milk-borne illnesses requires an understanding of the sorts of bacteria that could be transmitted by post-pasteurization contamination. [2]. In this study, we will use chemometric and dielectric spectroscopic methods to predict the qualitative characteristics of raw, unpasteurized milk. Milk that was collected at a temperature of 25°C and measured between 70 and 100% freshness. Physiochemical compositions including lipids, proteins, and water content are used to describe it, together with its dielectric properties in the 0.5 to 9 GHz frequency range. [3]. Chemical analysis of milk has traditionally been a laborious and time-consuming process that is also quite polluting. In this research, the quality of milk was evaluated using machine learning methods. We needed a new method to make it possible to determine the composition of milk on the spot in a fast, straightforward, and accurate manner. The article delves into the process of collecting feature data from several waves at once using a broadband infrared light source and a multichannel infrared spectral sensor [4]. Thus, in order to guarantee the quality of the milk, it has to be checked for the presence of all necessary components and any possible adulterants [5]. Here, pH, turbidity, and color are only a few of the factors that are calculated using sensors. The milk business needs the ability to provide the government with continuous data on milk quality as it makes milk packaging so that they can fight against illicit items like low-quality milk. As a result, there is an urgent need for a more precise way to quickly evaluate milk quality. One practical method that may be used to accomplish this objective is machine learning (ML). The field of artificial intelligence known as machine learning focuses on teaching computers to identify complicated patterns in data sets. This case employs the usage of two machine learning algorithms to evaluate milk quality. We utilize the milk quality dataset that is accessible

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

in the Kaggle repository for both the model's training and validation.

2.Literature Review

The Multiwavelength Gradient-Boosted Regression Tree was suggested by Sheng et al. [6] for the purpose of analyzing milk's protein and fat content. A technique for determining the wavelength intensity of milk was devised by Multiwavelength Spectral Sensor System, who used a multichannel spectral sensor. To measure how well the GBRT regression model worked, we looked at its explained variance regression score, mean square error, and coefficient of determination. In order to gather classification data, Brudzewski et al. [7] suggested using the SVM model. The object recognition and classification system is based on support vector machines (SVMs) with radial and linear kernels. In order to categorize milk, they used a system that relied on gas sensors based on oxides. Moharkar, Wasudeo et al. [8] Laser-Induced Detection and Quantification of Milk: A Proposed System. One typical method for detecting milk adulterants is laser-induced spectrometry. The Raspberry Pi is used for embedded data collection. Suggested by Kumar et al. [9] Ensemble machine learning is used in this study. It makes use of support vector machines (SVMs) and residual neural networks. Classification and regression tasks are often handled by support vector machines (SVMs), a kind of supervised learning technology; image recognition applications often make use of ResNets, a class of deep neural network designs. The goal of the study by Shobana et al. [10] is to create a system that recommends fruit consumption to visually impaired individuals. In order to extract features, the researchers used deep learning techniques including the Convolutional Neural Network and Visual Geometry Group 16. They also used RF, Light Gradient Boosting, and Logistic Regression as machine learning techniques for prediction. In their article on ELMs and ELMs based on kernels, Ruifang et al. [11] used. Their goal was to find out how these models fared in comparison to the popular BP neural network and SVM network. In order to evaluate water quality using predictive machine learning, Ghosh et al. (2023) set out on a thorough investigation. Their findings demonstrated that machine learning methods have great promise for improving water quality assessment and classification. Parameters including as pH, dissolved oxygen, BOD, and TDS were included in the dataset utilized for this purpose. The Random Forest model had an impressive 78.96% accuracy rate, making it the most accurate of the models they tested. The SVM model, on the other hand, was far behind, with an accuracy of only 68.29% [18]. To improve underwater picture dehazing, Alenezi et al. (2021) created a new Convolutional Neural Network (CNN) that combines a block-greedy algorithm. This technique improves both the local and global pixel values while also dealing with the attenuation of color channels. The method improves picture edges by using a one-of-a-kind Markov random field. Underwater photographs were found to be crisper, clearer, and more colorful when this approach was evaluated using measures like UIQE and UIQM, which showed that it was superior to previous methodologies [19]. With an emphasis on the rapid and substantial disruption it caused, Sharma et al. (2020) offered a thorough analysis of how COVID-19 affected global financial indices. According to the data, the world's markets lost more over US $6 trillion in a single week in February 2020, signaling a severe economic slump. They were able to shed light on how containment practices affected several financial measures via their multivariate research. This research highlights the possibility of using sophisticated algorithms for detection and analysis as well as the far-reaching consequences of the epidemic on economic activity [20].

3.Proposed Work & Methodology

The different phases of the proposed system are shown in the following block diagram.

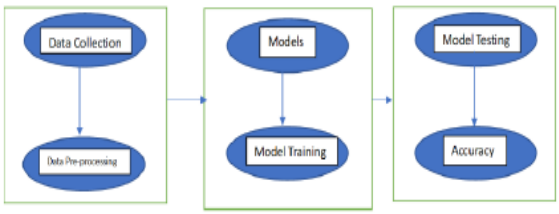

Fig1. System Architecture

3.1 Data Collection

The suggested system's data set Sourced from the Kaggle repository [12]. As can be seen in table1, this dataset is composed of seven distinct characteristics. Predicting the results of the milk analysis requires these characteristics. The milk's target grade, which is a kind of categorical data There are three distinct categories: low, moderate, and high. The dataset has 1059 rows and 8 columns totaling records. There is one numerical feature and seven category characteristics.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

Table 1. Categorical and Numerical data

Categorical Data	Numerical Data
Grade	pH
	Odor
	Temperature
	Taste
	Fat
	Color
	Turbidity

3.2 Data Pre-Processing

As a first step in pre-processing, we determined the missing value in the data. It turns out that every single feature has a complete set of data. After that, the process of label encoding is carried out. Because computers are completely unable of understanding the problem's attribute values, label encoding is used to convert them to integer categories. In this dataset, the 'Grade' feature is implemented using Label Encoding. The last step is to scale the feature values. This case makes use of the Min-Max scaling. According to equation 1, min-max scaling uses max values for scaling and min values for scaling. Normalizing features, using a method like Min-Max, should be done before fitting the model [13]. To deal with this possible problem, scaling is usually used [14].

$$m = (x - x_{min}) / (x_{max} - x_{min}) \quad (1)$$

where

x= feature value

x_{min}= minimum value of feature

x_{max}= maximum value of feature

Machine learning-based models outperform machinery-based systems when it comes to predicting milk quality. The suggested system may be connected to any device that can measure various milk quality characteristics in order to improve real-time milk quality evaluations. Reduced calibration expenses and improved accuracy with minimal calibration data sets Adapting to different work settings is easier. The food industry could utilize it to verify the milk's manufacturing specs.

Models (3.3)

Training data is used in the process of building models. There is 80% training data and 20% test data in the dataset. The model is being evaluated with the help of RF and SVM.

Section 3.3.1 Decision Tree

When training for a classification or regression problem, Random Forest generates a class or mean prediction for each of the decision trees it forms using unseen data. This ensemble learning approach is known as Random Forest. A "forest" of decision trees, usually trained using a "bagging" method, is what it constructs. Random forests are able to handle feature-rich data sets and rank the characteristics in order of significance for milk quality prediction. Some milk samples may not have all the necessary measurements in the actual world. The use of missing data does not affect the accuracy of predictions made by random forests. Decision trees may be problematic due to their tendency to overfit. But by using several trees and averaging their outputs, random forests are able to accomplish superior generalization. [11]. The supervised machine learning technique known as random forests is widely used for classification and regression problems, and it usually produces good results even when the hyperparameters are not changed. Reason being, it takes entropy into account when building a decision tree. "Gini" refers to the Gini impurity, while "entropy" denotes the increase in information. Equation 2 shows the formula for calculating the Gini index.

$$\text{Gini} = 1 - \sum_{i=1}^{c}(p_i)^2 \quad (2)$$

Where

p_i=proportion of data belongs to class c

3.3.2 Support Vector Machine (SVM)

A supervised machine learning approach, support vector machines (SVMs) are often used for regression and classification jobs. Finding the hyperplane that most effectively fits the data, or best classifies it in regression, is the main principle. SVM does this by making the most of the distance between the hyperplane and the support vectors, which are the data points closest to the two classes. SVM is well-suited for multi-parameter milk quality prediction because it can efficiently manage high-dimensional data. By using kernel functions like sigmoid, polynomials, and radial basis functions (RBF), support vector machines (SVMs) are able to deal with parameter relationships that are not linear. Carefully selected margins make SVMs less susceptible to overfitting. Improving the ability to summarize data that is not yet available is essential for making accurate forecasts of future milk samples. [11]

Fourth, Analyzing Performance

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

The purpose of this proposed study is to predict milk quality using RF and SVM. Various performance matrices, such the Performance Score in Table 2, are used to assess the classifiers' prediction efficiency.

Table 2. Performance Score

	Accuracy	Precision	Recall	F1 Score
RF	0.92	0.85	1.00	0.92
SVM	0.57	0.51	0.92	0.66

To calculate a classifier's accuracy, take the total number of predictions and divide it by the number of right predictions [15]. The accuracy with which a classifier separates the predicted total positives from the actual positives is known as its precision [16]. The classifier's recall is defined as the percentage of actual predictions that turn out to be true positives [17]. Using Precision and Recall, the F1 score evaluates the model's efficiency. Equations3,4,5, and 6 provide the formula for the aforementioned performance metrics.

$$\text{Accuracy} = \frac{J+K}{J+K+L+M}$$

$$\text{Recall} = \frac{J}{J+M}$$

$$\text{Precision} = \frac{J}{J+L}$$

$$\text{F1 Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (6)$$

Where
J = True positive
K= True negative
L = False positive
M = False negative

5.Conclusion

Applying support vector machines and RF forms the basis of the method for milk detection in the study. A semiconductor gas sensor array installed inside a measuring test chamber has been used to measure the Grade. Findings from numerical analyses of various milk production techniques and fat levels have shown the remarkable effectiveness of the proposed strategy. You can find out how much fat is in the milk even when you look at products manufactured by the same dairy. For modest training data sets, the RF application-based approach shown here performs well in terms of generalization. Machine learning-based models outperform machinery-based systems when it comes to predicting milk quality. The suggested system may be connected to any device that can measure various milk quality characteristics in order to improve real-time milk quality evaluations. Enhanced precision with a smaller calibration data collection and lower calibration expenses. Adapting to different work settings is easier. The food industry could utilize it to verify the milk's manufacturing specs.

References

1.Anderson, Melisa, et al. "The microbial content ofunexpired pasteurized milk from selected supermarketsin a developing country." Asian Pacific journal oftropical biomedicine 1.3 (2011): Volume 1, Issue 3,2011, Pages 205-211, ISSN 2221-1691,doi:10.1016/S2221-1691(11)60028-2.

2.Dhanashekar R, Akkinepalli S, Nellutla A. “Milk-borneinfections. An analysis of their potential effect on themilk industry”. Germs. 2012 Sep 1;2(3):101-9. doi:10.11599/germs.2012.1020. PMID: 24432270;PMCID: PMC3882853.

3.Wenchuan Guo, Xinhua Zhu, Hui Liu, Rong Yue,Shaojin Wang,"Effects of milk concentration andfreshness on microwave dielectric properties", Journalof Food Engineering, Volume 99, Issue 3,2010, Pages344-350,ISSN 0260-8774, doi:10.1016/j.jfoodeng.2010.03.015.

4.J. N. V. R. Swarup Kumar, D. N. V. S. L. S. Indira, K.Srinivas and M. N. Satish Kumar, "Quality Assessment and Grading of Milk using Sensors and Neural Networks," 2022 International Conference on Electronics and Renewable Systems (ICEARS),Tuticorin, India, 2022, pp. 1772-1776, doi:10.1109/ICEARS53579.2022.9752269

5.L. W. Moharkar and S. Patnaik, "Detection and Quantification of Milk Adulteration by Laser Induced Instumentation," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT),Bombay, India, 2019, pp. 1-5, doi:10.1109/I2CT45611.2019.9033883.

6.T. Sheng, S. Shi, Y. Zhu, D. Chen and S. Liu, "Analysis of Protein and Fat in Milk Using Multiwave length Gradient-Boosted Regression Tree," in IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-10, 2022, Art no. 2507810, doi:10.1109/TIM.2022.3165298

7.K. Brudzewski a, S. Osowski b, T. Markiewicz b ,“Classification of milk by means of an electronic noseand SVM neural network”. Received 30 June 2003,Revised 13 October 2003, Accepted 21 October 2003,Available online 30 December 2003.

8.L. W. Moharkar and S. Patnaik, "Detection and Quantification of Milk Adulteration by Laser Induced Instrumentation," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-5, doi:10.1109/I2CT45611.2019.9033883.

9.A. K. S, H. M. L, S. V. G. V., U. M.S, L. Kannagi and P. S. Bharathi, "A Novel and Effective Ensemble Machine Learning Model for Identifying Healthy and Rotten Fruits," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India,2023, pp. 1-7, doi:10.1109/ICECONF57129.2023.10083721.

10.S. G, Reethu, S. S and V. K, "Fruit Freshness Detecting System Using Deep Learning and Raspberry PI," 2022International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2022, pp. 1-7, doi:10.1109/ICSES55317.2022.9914056.

11.R. Zhang et al., "Prediction of Dairy Product Quality Risk Based on Extreme Learning Machine," 2018 2ndInternational Conference on Data Science and Business Analytics (ICDSBA), Changsha, 2018, pp. 448-456, doi: 10.1109/ICDSBA.2018.00090.

12.https://www.kaggle.com/datasets/prudhvignv/milk-grading

13.A. Deshpande, S. Deshpande and S. Dhande, "NIR Spectroscopy Based Milk Classification and Purity Prediction," 2021 IEEE Pune Section International Conference (PuneCon), Pune, India, 2021, pp. 1-5, doi:10.1109/PuneCon52575.2021.9686473.

14.Pires IM, Hussain F, M. Garcia N, Lameski P, Zdravevski E. Homogeneous Data Normalization and Deep Learning: A Case Study in Human ActivityClassification. *Future Internet*. 2020; 12(11):194.https://doi.org/10.3390/fi12110194.

15.Kumar, S., Neware, N., Jain, A., Swain, D., Singh, P.(2020). "Automatic Helmet Detection in Real-Timeand Surveillance Video". Advances in Intelligent Systems and Computing, vol 1101. Springer, Singapore. https://doi.org/10.1007/978-981-15-1884-3_5

16.Swain, D.; Mehta, U.; Bhatt, A.; Patel, H.; Patel, K.;Mehta, D.; Acharya, B.; Gerogiannis, V.C.; Kanavos, A.; Manika, S. A Robust Chronic Kidney Disease Classifier Using Machine Learning. Electronics 2023,12, 212. https://doi.org/10.3390/electronics12010212

17.Swain, D., Parmar, B., Shah, H., Gandhi, A., Pradhan, M.R., Kaur, H. & Acharya, B. (2022). Cardiovascular Disease Prediction using Various Machine Learning Algorithms. Journal of Computer Science, 18(10), 993-1004. https://doi.org/10.3844/jcssp.2022.993.1004

18.Ghosh, H., Tusher, M.A., Rahat, I.S., Khasim, S.,Mohanty, S.N. (2023). Water Quality Assessment Through Predictive Machine Learning. In: Intelligent Computing and Networking. IC-ICN 2023. Lecture Notes in Networks and Systems, vol 699. Springer,Singapore. https://doi.org/10.1007/978-981-99-3177-4_619.Alenezi, F.; Armghan, A.; Mohanty, S.N.; Jhaveri, R.H.; Tiwari, P. Block-Greedy and CNN Based Underwater Image Dehazing for Novel Depth Estimation and Optimal Ambient Light. Water 2021,13, 3470. https://doi.org/10.3390/w13233470

20.G. P. Rout and S. N. Mohanty, "A Hybrid Approach for Network Intrusion Detection," 2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, India, 2015, pp. 614-617, doi:10.1109/CSNT.2015.76.