EAI Endorsed Transactions
on Internet of Things

Research Article EAI.EU

Milk Quality Prediction Using Machine Learning

Drashti Bhavsar [1], Yash Jobanputra[2], Nirmal Keshari Swain[3], Debabrata Swain[4, *]

[1,2, 4] Department of Computer Engineering, School of Technology, Pandit Deendayal Energy University, Gandhinagar, India
[3] Information Technology Department, Vardhaman College of Engineering, Hyderabad, India

Abstract

Milk is the main dietary supply for every individual. High-quality milk shouldn't contain any adulterants. Dairy products are sold everywhere in society. Yet, the local milk vendors use a wide range of adulterants in their products, permanently altering the evaporated. Using milk that has gone bad can have serious health consequences. On October 18 of this year, the Food Safety and Standards Authority of India (FSSAI), the nation's top food safety authority, released the final result of the National Milk Safety and Quality Survey (NMSQS) and declared the milk readily available in India to be "mostly safe." According to an FSSAI survey, 68.4% of the milk in India is tainted. The quality of milk cannot be checked by any equipment or special system. Milk that has not been pasteurized has not been treated to get rid of harmful bacteria. Infected raw milk may contain Salmonella, Campylobacter, Cryptosporidium, E. coli, Listeria, Brucella, and other dangerous pathogens. These microorganisms pose a major risk to your family's health. Manually analyzing the various milk constituents can be very challenging when determining the quality of the milk. Analyzing and discovering with the aid of machine learning can help with this endeavor. Here a machine learning-based milk quality prediction system is developed. The proposed technology has shown 99.99% classification accuracy.

Keywords: Machine Learning, Milk Quality Prediction, Random Forest, Support Vector Machine, Label Encoding

Received on 12 September 2023, accepted on 19 November 2023, published on 29 November 2023

Copyright © 2023 D. Bhavsar *et al.*, licensed to EAI. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/eetiot.4501

*Corresponding author. Email: debabrata.swain7@yahoo.com

1. Introduction

Milk is considered as a most useful supplement in everyone's daily life. Hence better health it is advisable to consume good-quality milk. A perishable product is a milk. Tons of milk can deteriorate with just one gram of milk that is of low quality or structure, resulting in significant financial losses. In a very short period of time, millions of bacteria can grow in spoilt milk. [1] In this way, circumstances that jeopardize human health may arise if people ingest milk or dairy products. In the USA each year around 48 million medical cases come due to food contamination. Because improper maintenance of dairy products can pose a transmission risk for a variety of pathogens and cause outbreaks of brucellosis, listeriosis, tuberculosis, etc., developing nations like India confront more simultaneous concerns. According to an FSSAI survey, 68.4% of the milk in India is tainted. Understanding the types of germs that might be conveyed by post-pasteurization infection is essential in order to successfully control the spread of milk-borne diseases. [2]. Dielectric spectroscopy and chemometric techniques will be used in this project to forecast the qualitative properties of fresh, unpasteurized milk. Fresh milk, which was collected from 70 to 100% at 25°C. It is characterized using physiochemical compositions such as fatty chemicals, proteins, and proportion of water as well as for its dielectric characteristics in the range of 0.5 to 9 GHz [3]. The traditional chemical methods for figuring out milk's makeup are labor- and time-intensive, but also quite polluting. Machine learning techniques were utilized in this study to assess milk quality. In order to enable quick, easy, and on-the-spot determination of milk composition, a new approach had to be developed. This article explains how to use a multichannel infrared spectral sensor and a broadband infrared light source to simultaneously capture multi-wave length feature data [4]. Therefore, assure the milk's quality, it must be examined for the presence of required ingredients and any potential adulterants [5]. In this instance, sensors are utilized to calculate several parameters, including as pH, turbidity, and color. In order to combat illegal products like low-quality milk, the milk industry should be able to submit ongoing data on milk quality to the administration during the production of milk packages. Hence it becomes an ultimate need to assess the quality of milk in a short duration of time with higher accuracy. To perform this task Machine learning (ML) can be applied as a useful tool. Machine learning is a subset of Artificial intelligence that mainly deals with enabling a computer to recognize complex patterns using historical data. Here 2 machine learning is used to for examining the grade of milk. For the training and validation of the model, the milk quality dataset available in the Kaggle repository is used.

2. Literature Review

Sheng et al [6] proposed Multiwavelength Gradient-Boosted Regression Tree to Analysis of Protein and Fat in Milk. Multiwavelength Spectral Sensor System developed a method for calculating the milk's wavelength intensity using a multichannel spectral sensor. The coefficient of determination, mean square error, mean absolute error, and explained variance regression score was used to evaluate the effectiveness of the GBRT regression model. Brudzewski et al [7] has proposed the SVM model to Obtain data for the classification. The system for recognizing and categorizing objects has been implemented using SVM neural networks with linear and radial kernels. They used a system based on oxide-based gas sensors for the purpose of classifying milk. Wasudeo Moharkar et al [8] Proposed Laser-Induced Instrumentation to Detection and Quantification of Milk. Using laser-induced spectrometry, a few common milk adulterants can be found. Used for embedded data collecting is the Raspberry Pi. Kumar et al [9] proposed Support vector machines (SVM) and residual neural networks are employed in the research's ensemble machine learning technique. SVM, a supervised learning technique, is employed for classification and regression tasks, whereas ResNets, a class of deep neural network design, are frequently employed for image recognition applications. Shobana et al [10] have proposed in the paper aim to develop a fruit intake recommendation system for blind people. The researchers employed deep learning methods for feature extraction, such as Visual Geometry Group 16 and Convolutional Neural Network. Additionally, they employed machine learning methods like Logistic Regression, Light Gradient Boosting, and Random Forest (RF) for prediction. Ruifang et al [11] used in the paper Extreme Learning Machines and kernel-based Extreme Learning Machines. They compared these models' performance to that of the widely used BP neural network and SVM network. Ghosh et al. (2023) embarked on a comprehensive study to assess water quality through predictive machine learning. Their research underscored the potential of machine learning models in effectively assessing and classifying water quality. The dataset used for this purpose included parameters like pH, dissolved oxygen, BOD, and TDS. Among the various models they employed, the Random Forest model emerged as the most accurate, achieving a commendable accuracy rate of 78.96%. In contrast, the SVM model lagged behind, registering the lowest accuracy of 68.29%[18].

Alenezi et al. (2021) developed a novel Convolutional Neural Network (CNN) integrated with a block-greedy algorithm to enhance underwater image dehazing. The method addresses color channel attenuation and optimizes local and global pixel values. By employing a unique Markov random field, the approach refines image edges. Performance evaluations, using metrics like UCIQE and UIQM, demonstrated the superiority of this method over existing techniques, resulting in sharper, clearer, and more colorful underwater images[19].

Sharma et al. (2020) presented a comprehensive study on the impact of COVID-19 on global financial indicators, emphasizing its swift and significant disruption. The research highlighted the massive economic downturn, with global markets losing over US $6 trillion in a week in February 2020. Their multivariate analysis provided insights into the influence of containment policies on various financial metrics. The study underscores the profound effects of the pandemic on economic activities and the potential of using advanced algorithms for detection and analysis[20].

3. Proposed Work & Methodology

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

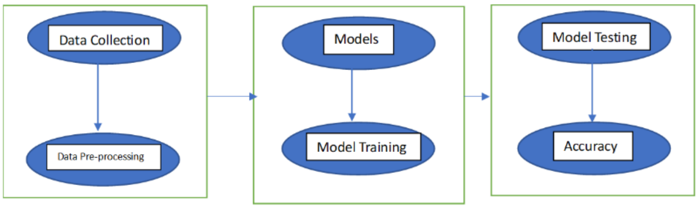

Fig1. System Architecture

3.1 Data Collection

The data set for the proposed system Collected from Kaggle repository[12]. This dataset consists of 7 independent features as shown in the table1. These parameters are used to predict analysis of the milk. Grade (Target) of the milk which is categorical data

Where Low (Bad) or Medium (Moderate) High are three different classes.The total number of records present in the dataset is 1059 rows, and 8 columns Out of all features, 7 are categorical and 1 is numeric.

Table 1. Categorical and Numerical data

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

3.2 Data Pre-Processing

In the first stage calculated the data's missing value in pre-processing. It is discovered that none of the features have a missing value. In the next step the Label encoding is performed. The value of the attribute in the problem cannot be understood by a computer at all, hence the values in this situation are transformed to category integer values using label encoding. Label Encoding is used for the 'Grade' feature in this dataset. Finally scalling of the feature values is performed. The Min-Max scaling is applied here. In Min-max scaling max value and min values are used for scale as shown in the equation 1. Before model fitting, firstly feature-wise normalization, such as Min-Max[13]. Scaling is typically employed to address this potential issue[14].

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

where
x= feature value
x_{min}= minimum value of feature
x_{max}= maximum value of feature

Using the Machine learning-based model the milk quality prediction can be done with more accuracy which is better than the machinery-based system. For a better assessment of milk quality in a real-time environment, the proposed system can be integrated with any device that can acquire the value of differ rent milk quality parameters.

Better accuracy at the relatively small size of a calibration data collection, cheaper calibration costs Ease of adaptation to various working environments. It might be used in the food business for checking the milk's production parameters.

3.3 Models

In model training, we are using training data. In the dataset, the training data is 80% and the testing data is 20%. We are using RF and SVM to evaluate the model.

3.3.1 Random Forest

Random Forest is an ensemble learning method that creates various decision trees during training and outputs a class (for classification tasks) or mean prediction (for regression tasks) for each tree on unseen data. The "forest" it builds is a collection of decision trees, typically trained using a "bagging" approach. Random forests can manage data sets with many features and determine the importance of each feature in predicting milk quality. In a real-world scenario, some milk samples may be missing certain measurements. Random forests can handle missing values and still produce accurate predictions. One of the challenges with decision trees is that they tend to overfit. However, random forests can achieve better generalization by using multiple trees and averaging their results. [11].

In regression and classification issues, random forests, a supervised machine learning method, are frequently employed and, most of the time, offer excellent results even without hyperparameter modification. Because of It constructs a decision tree and Entropy criteria consider.
"Gini" stands for the Gini impurity, while "entropy" stands for information gain. Gini index is calculated as shown in equation2.

$$\mathbf{Gini = 1} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{c}} (p_i)^{\mathbf{2}} \qquad \mathbf{(2)}$$

Where
p_i=proportion of data belongs to class c

3.3.2 Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm mainly used for classification and regression tasks. The core idea is to find the hyperplane that best classifies the data, or in the case of regression, fits the data most efficiently. SVM does this by maximizing the distance between the hyperplane and the closest data points in the two classes, called support vectors. SVM can effectively handle high-dimensional data and is therefore suitable for predicting milk quality based on multiple parameters. SVM can handle nonlinear relationships between parameters using kernel functions such as radial basis functions (RBF), polynomials, and sigmoid. SVMs are less prone to overfitting, especially when margins are chosen carefully. This means that data that is not yet visible can be better summarized, which is crucial for accurate predictions of new milk samples. [11]

4. Performance Analysis

In this proposed work RFand SVM are used for the milk quality prediction. The prediction effectiveness of the classifiers is determined using different performance matrices such as Performance Score shown in Table 2.

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

Accuracy of a classifier predicts the Number of correct predictions from the Total number of predictions [15]. Precision indicates how well a Classifier predicts True Positive from Total Positive Predicted [16]. Recall defines how well the classifier predicts true positives from Actual predicted [17]. The F1 score measures the efficiency of the model using the Precision and Recall. The formula for the above performance parameters shown in the equation 3, 4, 5 and 6.

$$\textbf{Accuracy} = \frac{J+K}{J+K+L+M} \qquad \textbf{(3)}$$

$$\textbf{Recall} = \frac{J}{J+M} \qquad \textbf{(4)}$$

$$\textbf{Precision} = \frac{J}{J+L} \qquad \textbf{(5)}$$

$$\textbf{F1 Score}= \frac{2*Precision*Recall}{Precision+Recall}$$

$$\textbf{(6)}$$

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

RF has more accuracy than SVM. RF algorithm is assemble-based. RF is a better learning algorithm because of more classifiers than SVM. Also better Decision making. RF mapping is in Information Gain. Information Gain shows how much pure classification. This helps in RF to burst accuracy, precision, recall, and F score.

5. Conclusion

The technique for identifying milk in the paper is based on the application of SVM and RF. The Grade has been measured using a semiconductor gas sensor array set inside a measuring test chamber. The outcomes of numerical studies identifying different milk production methods and fat contents have demonstrated the excellent efficacy of the suggested approach. Even within the family of milky goods made by a specific dairy, it can identify the milk's fat content. The proposed method, which is based on the RF application, has excellent generalization properties for relatively small training data sets. Using the Machine learning-based model the milk quality prediction can be done with more accuracy which is better than the machinery-based system. For a better assessment of milk quality in a real-time environment, the proposed system can be integrated with any device that can acquire the value of differ rent milk quality parameters. Better accuracy at the relatively small size of a calibration data collection, cheaper calibration costs. Ease of adaptation to various working environments. It might be used in the food business for checking the milk's production parameters.

References

1. Anderson, Melisa, et al. "The microbial content of unexpired pasteurized milk from selected supermarkets in a developing country." Asian Pacific journal of tropical 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-borne infections. An analysis of their potential effect on the milk 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 and freshness on microwave dielectric properties", Journal of Food Engineering, Volume 99, Issue 3,2010, Pages 344-350,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 Multiwavelength 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 nose and 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 Instumentation," 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," 2022 International 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 2nd International 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 Activity Classification. *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-Time and 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_6
19. 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.