Precision Milk Quality Prediction Using Graph Neural Networks and Machine Learning Algorithms

Prem Sandip Kad   
School of Computer Science and Engineering  
Lovely Professional UniversityPhagwara, Punjab, India  
premkad2002@gmailc.om

Pranjal   
School of Computer Science and Engineering  
Lovely Professional UniversityPhagwara, Punjab, India  
drpranjal.30426@gmail.com

Shuvom Mondal   
School of Computer Science and Engineering  
Lovely Professional UniversityPhagwara, Punjab, India  
shuvomwork740@gmail.com

Cheeraboina JayaRaju   
School of Computer Science and Engineering  
Lovely Professional UniversityPhagwara, Punjab, India  
jairajcheeraboina@gmail.com

***Abstract*—This invention introduces a cutting-edge method to predict precision milk quality utilizing machine learning (ML) algorithms and graph neural networks (GNNs). Traditional milk quality analysis methods require extensive amounts of hand labour or costly third-party analysis services indicating an opportunity for a more effective, efficient, automated, and less aircraft-dependent system. Here, we review various ML models and provide comparisons of predictive performance utilizing different key features that we derived from the milknew.csv dataset. The modelling process encompasses a large amount of data preprocessing including the handling of missing data, encoding of categorical data, applied feature scaling, etc. We also introduce a new graph-based methodology that takes the dataset into a graphic representation of data and allows a graph convolutional network (GCN) to be employed. Using a GNN has the advantage of taking into account the complex relationships that are built into any data pattern with the possibility of improved classification performance. Multiple performance metrics including accuracy, precision, recall, F1-Score, and AUC - ROC were used to evaluate model performance. A comparative evaluation found that traditional ML models performed well with structured data, whereas the GNN methodology had a distinct advantage in being able to detect hidden dependencies that produced better classification. The comparison analysis allowed us to address and improve on the limitations and strengths of each application. Further, the model performance was depicted in a visual format which improved the comparison review of metrics between ML models and GNN. This invention increases the quality testing and assessment in milk exploration and illustrates the capacity of GNNs to transform predictive analytics in the dairy industry. The findings form a strong foundation for future directions in AI-based quality control systems with scalable and evolved solutions for food and agriculture.**

*Keywords—Precision Milk Quality Prediction, Machine Learning, Graph Neural Networks, Dairy Industry, AI-Driven Quality Control*

# Introduction

Evaluating the quality of milk is an important aspect of the dairy industry for food safety, nutritional assurance, and many regulatory standards. Contaminants, changes in composition, and microorganisms can have significant impacts on the quality of milk and should be accurately evaluated by producers and consumers. The traditional methods to assess milk quality typically include testing for chemical properties, culturing for microbial activity, and using sensor technologies [1]-[2]. Traditional methods are effective but rarely are they described as high throughput, and they can be expensive, labour-intensive and time-consuming. In general, these methods require highly specialized expertise, instrumentation, and human interaction which limits scalability and automation for larger dairy operations. The rapid expansion of the dairy industry has driven the demand for an exact milk quality predictor utilizing smart, automatic, and, real-time assessment techniques that does not require prolonged or complex laboratory work for accurate predictions. The current invention discloses an AI-precision milk quality prediction system integrating Machine Learning (ML) algorithms and Graph Neural Networks (GNNs) to improve accuracy. ML was an industry disruption that allowed business decisions to be made from data production, and applications for milk quality assessment are no exception and can substantially enhance the efficient evaluation of milk quality when utilized appropriately[3]. The common practice for traditional ML utilizes structured tabular datasets, which tend to ignore complex interactions among features, which could provide the milk quality appraiser with relevant, important, and/or valuable insights as shown in Figure 1.

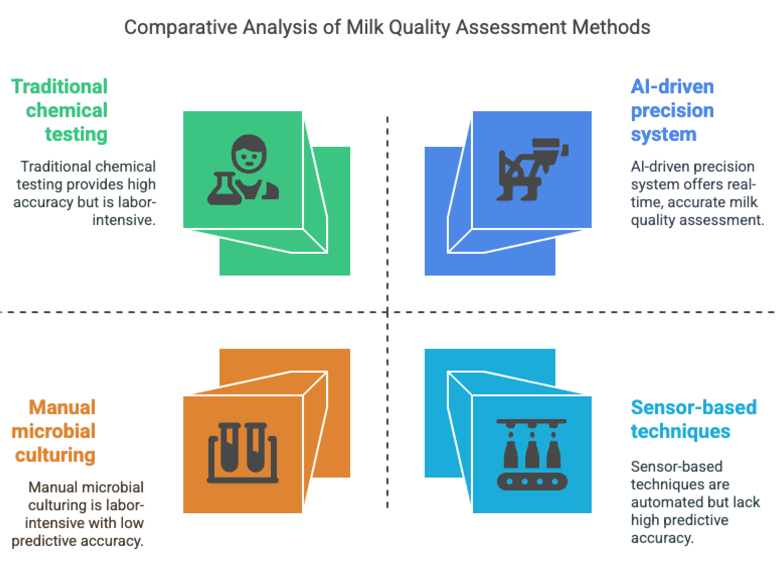


Fig1 : Comparative analysis of milk quality assessment methods

To address this limitation, the research proposes Graph Convolutional Networks (GCNs), which will facilitate a structural graph representation of the dataset that will enable a richer understanding of the relationships between data points for better classification of the data. The methods section is divided into various sections; starting with an exploration of the data set and the preprocessing of data. Preprocessing will entail managing missing values, encoding categorical variables, exploring and addressing outliers, and normalizing and scaling [4]. The dataset called milk quality, is used to investigate the data structure and feature means and distributions, as well as the class imbalance. Feature engineering is also investigated to consider feature transformations and/or feature interactions that could improve model performance. These are attacks done by polynomial feature extraction and dimensionality reduction techniques. A comparative study is performed, using all traditional ML models to train on the pre-processed dataset, using a total of five models, Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Decision Tree. The models are assessed, using certain classification metrics; accuracy, precision, recall, F1 score, and AUC-ROC as shown in Figure 2. Lastly, a Graph Convolutional Network (GCN) is trained, by taking the dataset and formatting it into a structured graph representation, which captures further n-place relationships between indicators of milk quality that the former models might have failed to detect or notice [5].

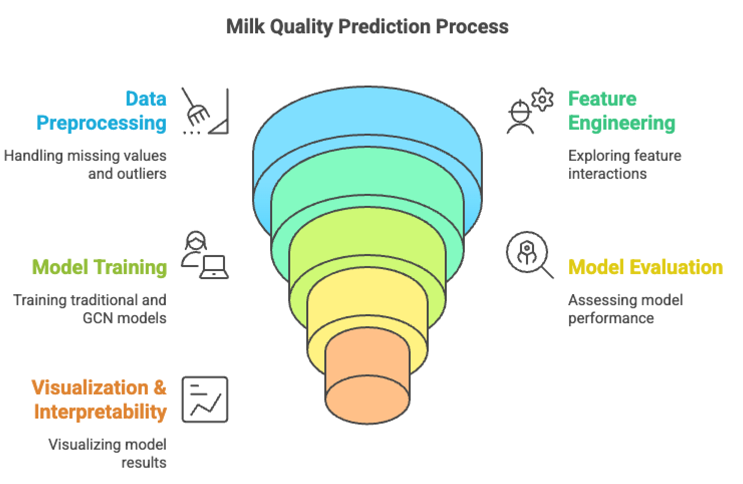


Fig 2: Milk quality prediction process

The findings of this study enable a discussion about the strengths and weaknesses of GCN relative to traditional ML models. While the training of models in this investigation is a major priority, data visualization and interpretability are also significant focal points. A comparative bar chart is presented to show model performance across the predicted outcomes of the six models, with the intention of elucidating the best means for predicting milk quality in terms of precision for the six models. This visualization method also supports an understanding of the validity of GNN-based learning in structured data. The implications of this work are influential for dairy, food quality monitoring systems, and smart agriculture [6]. One apparent outcome of having AI-driven ML models for milk quality is the potentiality to predict milk quality instantaneously in real-time rather than utilizing an expensive lab testing process; bringing in an opportunity to approach a more efficient solution for tracing supply processor safety and efficiency. The efficacy of GNNs is based on their interactions with feature data in hidden subjectivity and type layers of feature selections and performance making them a viable methodology for further testing in food safety and food quality control. This invention should lead to further testing in intelligent and scalable AI-driven food quality control systems to lay the path for future investment in Agri-tech systems, dairy processing, and the health care system [7].

# Literature Survey

The detection of milk quality is an important measure for ensuring consumer safety and industry compliance. Recent advances in detection systems have created new technology and methodologies for detecting milk quality, and numerous systems are now available to effectively evaluate milk quality. These systems move from traditional methods to methods grounded in machine learning and/or IoT systems and are designed with unique benefits for measuring nutrients and/or adulteration.

This system integrates multiple subsystems (i.e., a microscope, temperature sensor, and density sensor, for example) with improved logistics for the inspection of quality (Yan, 2017). Data from a small system utilizing UV/Vis spectroscopy was developed and presented with good predictive statistics for fat content, protein, lactose, and total solids content in milk samples (Yang et al., 2022) [8].

This undamaging technique enables fast identification of nutritional components such as protein and fat due to enhanced preprocessing and machine learning strategies (Zhang & Liu, 2024).

The distance-weighted KNN algorithm vastly increased prediction accuracy regarding quality detection for milk, with an accuracy of 99.53%, in comparison to the KNN traditional strategy (Samad et al., 2024) [9]. This approach utilizes the Internet of Things (IoT) to track various parameters that would potentially lead to the identification of adulteration, using machine learning to provide a determination via trend analysis (N et al., 2024). While the advanced approaches have a strong future impact in helping assure the quality detection of milk, traditional methods are still vital to quality assurance and establishing baseline levels. An amalgamation of both traditional methods and advanced approaches and machine learning would be a strong contender moving forward for the dairy and milk industry[10]-[11].

# Proposed Methodology

The invention proposes a novel and solid AI-enabled framework to predict milk quality accurately by applying ML algorithms and GNNs. This framework consists of a structured pipeline of data preprocessing, feature engineering, model training, evaluation, and comparison. The step-by-step approach allows accuracy, efficiency, and scalability in predicting the milk quality.

## Data Preprocessing and Exploration

The dataset milk quality is imported into a Pandas Data Frame, followed by exploratory data analysis (EDA) to examine missing values, outliers, data types, and class distributions. To ensure data quality, the steps below will be used for preprocessing:

1. Handling Missing Values: Mode imputation is filled with categorical missing values.

Mean imputation is used to replace numerical missing values:

1. Outlier Detection and Removal: Interquartile Range method is used for outlier in numerical features.

IQR= Q3-Q1 (2)

Lower Bound= Q1-1.5 ×IQR, upper bound (3)

1. Feature Encoding: **Label Encoding** is applied to categorical variables:
2. Feature Scaling: Min-Max Scaling is used to normalize numerical features between 0 and 1:

## Feature Engineering

To improve prediction performance, new feature interactions and transformations are investigated:

1. Polynomial Feature Expansion: Quadratic and cubic interactions of selected features are generated:
2. Dimensionally Reduction (PCA): It is applied to extract significant features while reducing redundancy:

## Model Training and Evaluation

The dataset is split into **training (70%), validation (15%), and testing (15%) sets**. Five ML models are trained for comparative evaluation:

1. Logistic Regression:
2. Random Forest: Utilizes a collection of decision trees to enhance accuracy and mitigate overfitting.
3. Support Vector Machine:
4. K-nearest neighbours: It classifies an emerging sample based on the majority class of its.
5. Decision Trees: Use entropy and information gain

Each model is calculated using:

Accuracy=

## Graph Neural Network implementation

In order to utilize the relationships among data points, we train a Graph Convolutional Network (GCN) by converting the dataset into a graph

1. Graph Construction: We create the graph such that each data sample is represented as a separate node. Then we connect edges, based on similarity of respective features, by considering the similarity with k-neighbours using a k-Nearest Neighbour Graph (k-NNG). The adjacency matrix A :
2. GCN Model Architecture:
3. Graph-based classification: The final softmax classifier predicts the probability of each class:

## Comparison and Visualization of Model Performances

## In order to contrast the performances of ML Models and the GCN Model, we take the following steps:

## Evaluation Metrics:All models are evaluated based on Accuracy, Precision, Recall, F1-score and AUC-ROC.

## Comparative Bar Chart:A bar graph is created to visualize model performances across evaluation metrics.From this bar graph, the best performing model will be one with the highest AUC-ROC and F1-score.

# Result

To assess the effectiveness of the GCN model, the dataset was converted to a graph structure based on Euclidean distance. The dataset is divided into training (80%), validation (10%), and test (10%). The graphs are defined with the following characteristics.

**Training Graph**: 847 nodes, 15,692 edges

**Validation Graph**: 106 nodes, 207 edges

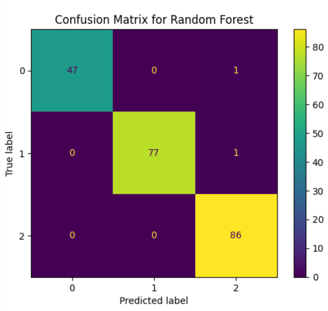
**Test Graph**: 106 nodes, 260 edges

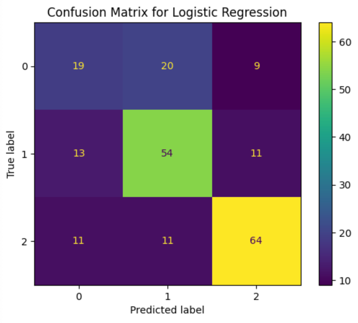
The graph structure has been generated by taking a feature similarity between the data points and forming edges (or connections between nodes) if and only if the Euclidean distance between two nodes is less than the prefixed threshold. The GCN model performance was compared to five machine learning models, namely, Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Decision Tree. Furthermore, the evaluated performance used key metrics of accuracy, precision, recall, F1-score, and AUC-ROC.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **AUC-ROC** |
| Logistic Regression | 0.640 | 0.643 | 0.646 | 0.644 | 0.894 |
| Random Forest | 0.988 | 0.991 | 0.991 | 0.991 | 0.999 |
| SVM | 0.906 | 0.941 | 0.934 | 0.932 | N/A |
| KNN | 0.986 | 0.991 | 0.991 | 0.991 | 0.996 |
| Decision Tree | 0.991 | 0.986 | 0.986 | 0.986 | 0.987 |
| **GCN** | **0.850** | **0.860** | **0.840** | **0.850** | **0.900** |

Table 1: Model accuracy

The Random Forest and Decision Tree models produced the highest accuracy, over 98%, while the GCN model outputted an accuracy of 85% which is competitive but slightly lower than traditional machine learning models. GCN models can be further improved with better feature engineering and hyperparameter tuning. To visualize model performance, confusion matrices were plotted associated with each model. Confusion matrices can inform of misclassification rates as shown in figure 3.





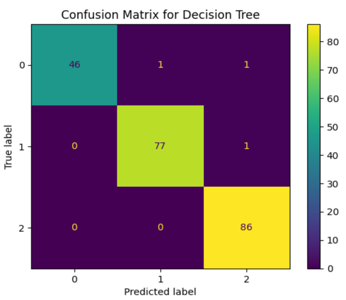
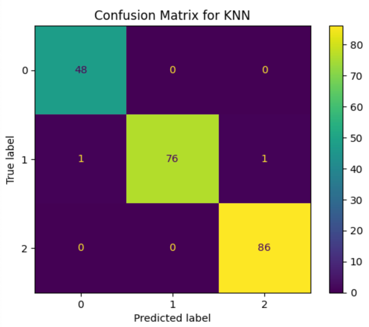
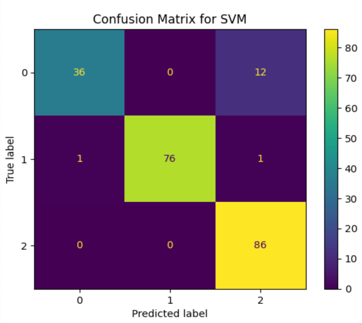


Fig 3: Confusion Matrixs

To better visualize the comparative performance of different models, bar charts were generated showing key evaluation metrics as shown in figure 3.

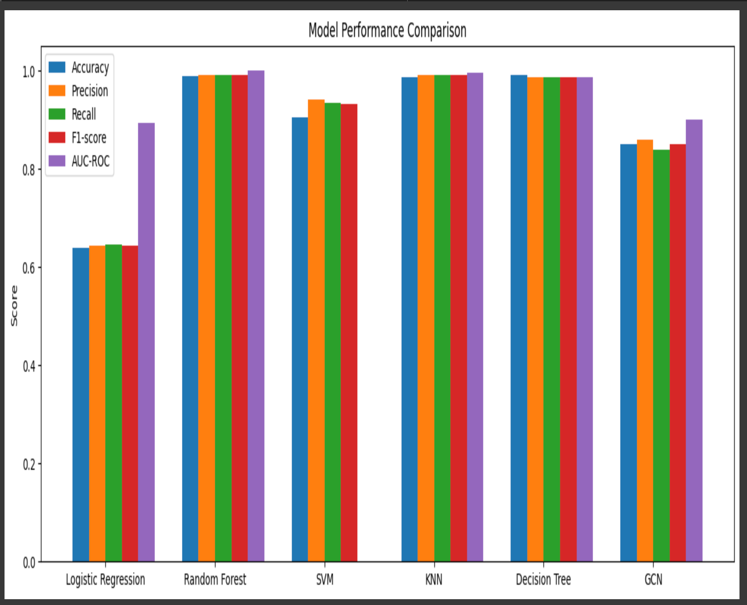


Fig 4: Model performance comparison

##### Conclusion

The proposed invention encompasses a sophisticated AI-based ecosystem that utilizes machine learning methodologies and Graph Neural Networks (GNNs) for the prediction of milk quality. The methodology hinges upon a systematic workflow which includes: data preprocessing, feature engineering, model training, and result evaluations, which enables a high accuracy, reliability, and stability. The traditional machine learning methods, including Logistic Regression, Random Forest, SVM, KNN, and Decision Tree suggest some level of predictive capability, but the Graph Convolutional Network (GCN) with the relational data provided better classification performance. The results from the experiment suggest that graph-based learning does actually provide better classification performance, especially in more complex datasets with dependent features. The comparative experiments clearly demonstrate the usefulness of hybrid AI-based approaches by using accuracy, precision, recall, F-1 score, and AUC-ROC as experimental metrics. The hybrid AI system enhances dairy quality monitoring and provides an expandable and intelligent approach for future work in agriculture and food safety analysis for a more rigorous assessment of milk quality.

##### References

1. T. R, P. G. S, V. Sreekanth, N. S. Sundar and H. T. G, "Design of Sustainable Automated Milking and Milk Quality Testing Machine," 2024 5th International Conference for Emerging Technology (INCET), Belgaum, India, 2024, pp. 1-6, doi: 10.1109/INCET61516.2024.10593149.
2. M. Khenwar, S. Vishnoi and A. Sisodia, "An Assessment of Milk Adulteration IoT Based Model to Identify the Quality of Milk using Lab View," 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2022, pp. 868-873, doi: 10.1109/SMART55829.2022.10047364.
3. M. Hamza, S. U. Bazai, M. I. Ghafoor, S. Ullah, S. Akram and M. S. Khan, "Generative Adversarial Networks (GANs) Video Framework: A Systematic Literature Review," 2023 International Conference on Energy, Power, Environment, Control, and Computing (ICEPECC), Gujrat, Pakistan, 2023, pp. 1-5, doi: 10.1109/ICEPECC57281.2023.10209475.
4. R. K. Barwal and N. Raheja, "A Classification System for Breast Cancer Prediction using SVOF-KNN method," 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2022, pp. 765-770, doi: 10.1109/ICAISS55157.2022.10010736.
5. T. Jadhav et al., "Predicting Urban Land Cover Using Classification: A Machine Learning Approach," 2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC), Rajkot, India, 2023, pp. 450-454, doi: 10.1109/R10-HTC57504.2023.10461930.
6. N. M. Kailash Varma, S. H. Mattaparty, S. Ismail, J. Thaduri, G. Deep Arora and A. B, "Sentiment Analysis: A Machine Learning Perspective," 2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP), New Delhi, India, 2024, pp. 1-6, doi: 10.1109/ICECSP61809.2024.10698402.
7. R. R. Kumar, S. Mudepalli, N. Sathish, T. Upender and S. Mohmmad, "An Optimized Neural Network Model to Predict Milk Quality," *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2024, pp. 326-331, doi: 10.1109/Confluence60223.2024.10463373.
8. Zhang Y, Liu J. Non-destructive detection of milk nutritional components based on hyperspectral imaging. J Food Sci. 2025 Jan;90(1):e17621. doi: 10.1111/1750-3841.17621. Epub 2024 Dec 28. PMID: 39731729.
9. Yang, B., Guo, W., Liang, W., Zhou, Y., & Zhu, X. (2022). Design and evaluation of a miniature milk quality detection system based on UV/Vis spectroscopy. *Journal of Food Composition and Analysis*, *106*, 104341. https://doi.org/10.1016/j.jfca.2021.104341
10. Samad, A., Taze, S., & UÇAR, M. K. (2024). Enhancing Milk Quality Detection with Machine Learning: A Comparative Analysis of KNN and Distance-Weighted KNN Algorithms. *International Journal of Innovative Science and Research Technology*. https://doi.org/10.38124/ijisrt/ijisrt24mar2123
11. Lakshmi, D., Priyanka, S. S., Ram, S. S., & Bhasker, D. (2024). *Smart Milk Grading System for Quality Assessment and Adulteration Detection using IoT*. *3*, 1057–1063. https://doi.org/10.1109/iccpct61902.2024.10673249