

Milk Quality Estimation Using Machine Learning Algorithms

Doddi Mohan Sai Vamsi

Department of MCA

AQJ Center for PG Studies

Andhra University, Visakhapatnam, Andhra Pradesh, India

Email: vamsih2385@gmail.com

Guide: Mrs. M. Tulasi

Abstract

Milk quality assessment is essential for ensuring food safety, maintaining consumer health, and improving the efficiency of dairy production. Conventional laboratory-based quality testing methods provide reliable results but require specialized equipment, skilled personnel, and significant processing time, making them unsuitable for rapid quality assessment in dairy farms and milk collection centers. Recent advancements in machine learning have enabled the development of intelligent systems capable of predicting milk quality using measurable physicochemical characteristics.

This research proposes a machine learning-based milk quality estimation system that classifies milk samples into **High**, **Medium**, or **Low** quality categories. The prediction is performed using seven physicochemical parameters, namely **pH**, **temperature**, **taste**, **odor**, **fat content**, **turbidity**, and **colour**. Four supervised machine learning algorithms—**Extra Trees Classifier**, **Random Forest Classifier**, **AdaBoost Classifier**, and **Artificial Neural Network (ANN)**—were implemented and comparatively evaluated to identify the most effective prediction model. Experimental evaluation demonstrated that the **Extra Trees Classifier** achieved the highest classification accuracy of **95.00%**, followed by Random Forest (**94.58%**), AdaBoost (**86.25%**), and ANN (**73.75%**). Based on this comparative analysis, the Extra Trees Classifier was selected for deployment.

To improve usability, the trained model was integrated into a **Flask-based web application** that enables users to enter milk quality parameters through a graphical interface and instantly obtain prediction results. The application also stores prediction history in an SQLite database and provides performance visualization through model comparison graphs and analytical dashboards. The proposed system offers a rapid, cost-effective, and user-friendly solution for preliminary milk quality assessment and demonstrates the practical application of ensemble machine learning techniques in dairy quality management.

Keywords— Milk Quality Estimation, Machine Learning, Extra Trees Classifier, Random Forest Classifier, AdaBoost Classifier, Artificial Neural Network, Flask, Food Quality Prediction.

II. LITERATURE REVIEW

Several researchers have investigated the application of machine learning techniques for food quality assessment and agricultural decision support. Traditional milk quality evaluation primarily relies on chemical and microbiological laboratory analysis, which provides accurate results but is time-consuming, costly, and dependent on specialized laboratory facilities. These limitations have encouraged the adoption of intelligent data-driven approaches for rapid quality prediction.

Ensemble learning algorithms such as **Random Forest** and **Extra Trees Classifier** have demonstrated strong performance in classification tasks because they combine multiple decision trees to improve prediction accuracy while reducing overfitting. Similarly, **AdaBoost** has been widely applied to enhance classification performance by combining weak learners into a stronger predictive model. **Artificial Neural Networks (ANNs)** have also been successfully used for nonlinear prediction problems due to their ability to learn complex relationships among input variables.

Previous studies on milk quality prediction have generally focused on physicochemical parameters including **pH, temperature, fat content, odor, taste, turbidity, and colour**. These parameters provide valuable information regarding milk freshness and quality and have been used to develop predictive models with promising classification performance. However, many existing studies evaluate only one or two machine learning algorithms or remain limited to offline experimental environments without providing a practical deployment platform.

The proposed work extends previous research by performing a comparative evaluation of **Extra Trees Classifier, Random Forest Classifier, AdaBoost Classifier, and Artificial Neural Network (ANN)** using the same dataset and evaluation strategy. Furthermore, the best-performing model is integrated into a **Flask-based web application** that enables real-time prediction, prediction history management, and graphical performance analysis. This combination of comparative machine learning analysis and web deployment provides a practical and user-friendly solution for rapid milk quality estimation.

III. PROPOSED METHODOLOGY

The proposed system estimates milk quality using supervised machine learning techniques. The workflow consists of data preprocessing, model training, model evaluation, model selection, and deployment through a Flask-based web application. Figure 1 illustrates the overall workflow of the proposed system.

Figure 1. Proposed System Workflow

(Insert your System Architecture or Workflow Diagram here.)

The methodology is divided into five major stages:

A. Data Collection and Preprocessing

The dataset contains milk samples described by seven physicochemical parameters that influence milk quality:

- pH
- Temperature
- Taste
- Odor
- Fat Content
- Turbidity
- Colour

The target variable classifies milk quality into three categories:

- High
- Medium
- Low

Before model training, the dataset is preprocessed by loading the data, checking for missing values, separating input features and target labels, and splitting the dataset into training and testing sets. These preprocessing steps improve data consistency and ensure reliable model evaluation.

B. Machine Learning Model Training

Four supervised machine learning algorithms were implemented and trained using the same dataset:

- Extra Trees Classifier

- Random Forest Classifier
- AdaBoost Classifier
- Artificial Neural Network (ANN)

Each model was trained using identical preprocessing and evaluation procedures to ensure a fair comparison. After training, the models were tested on unseen data to evaluate their classification performance.

C. Model Evaluation

The trained models were evaluated using commonly accepted classification metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

The comparative analysis identified the **Extra Trees Classifier** as the best-performing algorithm with an accuracy of **95.00%**. Consequently, this model was selected for deployment in the prediction system.

D. Model Deployment

The selected model was saved using the **Joblib** library and integrated into a **Flask-based web application**. Users can enter milk quality parameters through a web interface, and the application predicts the milk quality category in real time. Each prediction is stored in an SQLite database, enabling prediction history management and further analysis.

IV. MACHINE LEARNING ALGORITHMS

To identify the most suitable model for milk quality estimation, four supervised machine learning algorithms were implemented and evaluated using the same dataset and preprocessing strategy. Each algorithm was trained under identical conditions to ensure a fair comparison. The models were evaluated using Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.

A. Extra Trees Classifier (ETC)

The **Extra Trees Classifier (ETC)** is an ensemble learning algorithm that constructs multiple decision trees using randomly selected features and randomly generated split points. Unlike conventional decision tree algorithms, ETC introduces additional randomness during tree construction, reducing overfitting and improving generalization.

In this research, the Extra Trees Classifier achieved the highest prediction accuracy of **95.00%** and demonstrated superior classification performance compared with the other evaluated algorithms. Owing to its high accuracy and robustness, the trained ETC model was selected for deployment in the Flask-based prediction system.

B. Random Forest Classifier (RFC)

The **Random Forest Classifier** is an ensemble technique that combines multiple decision trees generated from bootstrap samples of the training dataset. The final prediction is obtained through majority voting among individual trees, improving prediction stability and reducing variance.

The Random Forest model achieved an accuracy of **94.58%**, indicating excellent classification performance. Although its performance was slightly lower than that of the Extra Trees Classifier, it provided reliable and consistent predictions for the milk quality dataset.

C. AdaBoost Classifier (ABC)

The **AdaBoost (Adaptive Boosting) Classifier** improves predictive performance by sequentially combining multiple weak learners into a stronger classifier. During training, higher importance is assigned to previously misclassified samples, allowing the algorithm to focus on difficult instances.

The AdaBoost model achieved an overall accuracy of **86.25%**. While its performance was satisfactory, it was lower than the ensemble tree-based models implemented in this study.

D. Artificial Neural Network (ANN)

The **Artificial Neural Network (ANN)** is a computational model inspired by the biological neural network of the human brain. It consists of interconnected layers of artificial neurons that learn complex nonlinear relationships through iterative training.

In the proposed system, ANN achieved an accuracy of **73.75%**. Although the model successfully learned patterns from the dataset, its performance was comparatively lower than the ensemble learning algorithms for the available structured data.

Set 6

Excellent! This is **SET 6**. This section is designed to occupy **approximately two journal pages**, including **only four figures**, keeping the overall paper within the **7–8 page** target.

Important: Insert **your actual project figures** where indicated.

V. EXPERIMENTAL RESULTS AND DISCUSSION

The performance of the proposed milk quality estimation system was evaluated by comparing four supervised machine learning algorithms using the same training and testing dataset. The evaluation focused on classification performance and model reliability using standard metrics including **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **Confusion Matrix**. All experiments were conducted using Python with the Scikit-learn library, and the selected model was deployed through a Flask web application.

A. Performance Evaluation

The prediction accuracy of each implemented algorithm is summarized in **Table III**.

Table III. Performance Comparison of Machine Learning Algorithms

Algorithm	Accuracy (%) Performance	
Extra Trees Classifier	95.00	Best
Random Forest Classifier	94.58	Excellent
AdaBoost Classifier	86.25	Good

Algorithm	Accuracy (%)	Performance
Artificial Neural Network	73.75	Moderate

The results demonstrate that the **Extra Trees Classifier** achieved the highest classification accuracy of **95.00%**, making it the most suitable algorithm for milk quality estimation. Random Forest also performed well with an accuracy of **94.58%**, while AdaBoost and ANN achieved comparatively lower performance.

Figure 2. Accuracy Comparison of Machine Learning Algorithms

Insert your "Accuracy Comparison" graph here.

Discussion:

The accuracy comparison graph illustrates that ensemble tree-based algorithms outperform the remaining classifiers for the given dataset. The Extra Trees Classifier consistently achieved the highest accuracy due to its randomized tree construction, which reduces overfitting and improves generalization.

B. Confusion Matrix Analysis

The confusion matrix provides a detailed view of the classification results by comparing the predicted classes with the actual class labels.

Figure 3. Confusion Matrix of the Extra Trees Classifier

Insert your Confusion Matrix here.

Discussion:

The confusion matrix indicates that the majority of milk samples were correctly classified into their respective quality categories. Only a small number of samples were misclassified, demonstrating the robustness and reliability of the selected prediction model.

C. Feature Importance Analysis

Understanding the contribution of input features improves model interpretability. The Extra Trees Classifier provides feature importance scores that indicate the relative influence of each physicochemical parameter on the prediction process.

Figure 4. Feature Importance Analysis

Insert your Feature Importance graph here.

Discussion:

The analysis shows that parameters such as **pH**, **temperature**, **odor**, and **fat content** contribute significantly to milk quality prediction. These features play an important role in distinguishing between High-, Medium-, and Low-quality milk samples.

D. Flask Web Application

The trained Extra Trees model was deployed using the Flask framework to enable real-time prediction through a web browser. The application provides an intuitive interface where users enter the seven milk quality parameters and receive an immediate prediction.

The web application includes the following functionalities:

- User-friendly prediction interface.
- Real-time milk quality estimation.
- Storage of prediction history using SQLite.
- Graphical performance dashboard.
- Model comparison and analysis.

Figure 5. Flask-Based Prediction Interface

Insert your Prediction Page or Result Page screenshot here.

Discussion:

The web interface simplifies interaction with the prediction model and enables users without programming knowledge to estimate milk quality efficiently. Integration with SQLite allows prediction records to be stored for future analysis and reporting.

E. Discussion

The experimental evaluation confirms that ensemble learning algorithms are highly effective for milk quality estimation using physicochemical parameters. Among the evaluated models, the **Extra Trees Classifier** demonstrated the best balance of prediction accuracy, computational efficiency, and robustness. Random Forest also achieved strong performance, indicating that tree-based ensemble methods are well suited to this structured classification problem.

AdaBoost produced acceptable results but showed lower classification performance compared with the ensemble tree methods. The Artificial Neural Network achieved the lowest accuracy in this study, suggesting that the available dataset and feature set are more effectively modeled using decision tree ensembles than neural network architectures.

The deployment of the selected model through a Flask web application demonstrates the practical applicability of the proposed system. Real-time prediction, prediction history storage, and graphical analysis make the application suitable for educational purposes and preliminary quality assessment in dairy environments.

Overall, the experimental results validate the effectiveness of the proposed methodology and support the selection of the **Extra Trees Classifier** as the final prediction model for the milk quality estimation system.

VI. CONCLUSION AND FUTURE SCOPE

A. Conclusion

This research presented a machine learning-based approach for estimating milk quality using physicochemical characteristics of milk. The proposed system utilized seven input parameters—**pH, temperature, taste, odor, fat content, turbidity, and colour**—to classify milk samples into **High, Medium, and Low** quality categories. Four supervised machine learning algorithms, namely **Extra Trees Classifier, Random Forest Classifier, AdaBoost Classifier, and Artificial Neural Network (ANN)**, were implemented and comparatively evaluated.

The experimental results demonstrated that the **Extra Trees Classifier** achieved the highest prediction accuracy of **95.00%**, outperforming Random Forest (**94.58%**), AdaBoost (**86.25%**), and ANN (**73.75%**). Based on this comparative analysis, the Extra Trees model was selected as the final prediction model and successfully deployed in a Flask-based web application.

The developed application provides real-time milk quality prediction through a simple graphical interface while maintaining prediction history using SQLite. The combination of machine learning and web deployment offers a practical, efficient, and cost-effective solution for preliminary milk quality assessment. Although laboratory testing remains essential for official certification, the proposed system can serve as an effective decision-support tool for dairy farmers, milk collection centers, educational institutions, and researchers.

B. Future Scope

The proposed system can be further enhanced by incorporating additional technologies and larger datasets. Future improvements include:

- Integration of IoT-based sensors for automatic collection of milk quality parameters.
- Development of Android and iOS mobile applications for portable prediction.
- Evaluation using larger datasets collected from different geographical regions.
- Deployment on cloud platforms for remote access and scalability.
- Integration with dairy management systems for continuous quality monitoring.
- Investigation of advanced deep learning techniques and explainable AI methods to improve prediction accuracy and interpretability.

These enhancements would increase the applicability of the system in commercial dairy environments and support intelligent decision-making in food quality management.

Acknowledgement

The author expresses sincere gratitude to **Mrs. M. Tulasi**, Project Guide, Department of MCA, AQJ Center for PG Studies, Andhra University, Visakhapatnam, for her valuable guidance, continuous encouragement, and technical support throughout this research work.

The author also thanks the faculty members of the Department of MCA, AQJ Center for PG Studies, Andhra University, for providing the necessary academic support and facilities required to complete this project successfully.