Qualitative Validation for Multimodal Deep Learning in Milk Quality Classification

Puteri Nur Farzanah Faghira Kamarudin1, a), Nik Mohd Zarifie Hashim1, b) and Noor Ziela Abd Rahman2, 3, c)

1Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.

2Faculty of Engineering and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia.

3Centre for Manufacturing and Environmental Sustainability Centre, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia.

c) Corresponding author: ziela.abdrahman@mmu.edu.my

a) m122310065@student.utem.edu.my   
b) nikzarifie@utem.edu.my

**Abstract.** This paper explores the effectiveness of intermediate fusion multimodal deep learning model in classifying milk quality by integrating both visual and numerical data. A qualitative analysis was conducted using an online questionnaire, where 93 participants classified milk samples based on visual cues alone and then with the addition of numerical data such as sample time, pH level, temperature, storage conditions, exposure status, and odor. Results showed a significant improvement in human classification accuracy when both data types were provided. In parallel, a custom intermediate fusion model was developed by combining features from a convolutional neural network (CNN) and a dense neural network. This model gave the best testing accuracy of 98.87%, displaying the top performing quality at 25 epochs and a batch size of 128. Finding from both human and machine analysis shows that multimodal inputs result in more accurate and reliable classification than unimodal approaches. This paper brings out the possibility of multimodal deep learning for real-world food quality monitoring and provides future smart dairy inspection systems with foundation.

# introduction

It is clear that milk as a perishable product loses its value if it is not stored correctly, this explains the rapid rate of spoilage in milk. The safety and quality assurance in milk is not only important for the public health but also for confidence building within the dairy supply chain. Assessment of milk quality has, since inception, been done by physically examining it. This process involves sensory evaluation, and chemical tests are also conducted on the milk that is used today. However, these methods are to some extent subjective, lengthy or require large laboratory facilities hence making them less feasible for automated solutions or scalable quality monitoring systems, so they need a new solution to monitor such problems efficiently. Over the past few years deep learning techniques have been developed as one of the powerful weapons against foods quality assessment automation. Unimodal deep learning approaches where they process data received from single sources like image or numerical readings, show quite good results. However, predictors based on individual spoilage features may be inaccurate due to complex conjunction of quantitative changes affecting combination change rate namely discoloration development stage, curdling changes development speed. Relying solely on one type of data may lead to misclassifications, especially in borderline cases where visual or numerical cues alone may not be conclusive. To address this limitation, multimodal deep learning integrates multiple data types to leverage complementary information and improve classification performance. This approach not only enhances model accuracy but also better mimics how humans assess food quality which is by combining what they see with contextual cues such as storage conditions and smell. In this paper, we present a qualitative validation of multimodal deep learning using a human-centered study conducted via Google Forms. The study consisted of three sections. The first section collected respondent’s general information such as gender, age, occupation, and milk consumption frequency. The second section asked respondents to classify nine milk samples using visual data only. The third section presented the same samples, but with both visual and numerical data, including sample time, pH value, temperature, storage condition, exposure status, and odor. Respondents were instructed to classify each sample as either ‘Good’, ‘Spoiling’, or ‘Spoiled’.

The purpose of this study is as follows.

* **Human VS Machine:** To analyze how human perception varies when only visual data is available versus when both visual and numerical data are presented.
* **Multimodal Deep Learning:** To support the hypothesis that multimodal deep learning outperforms unimodal methods in milk quality classification.

# literature review

Thanks to the progress in machine learning and computer vision, quality control of food has seen much improvement through monitoring systems. In the context of dairy products, particularly milk, several studies have been conducted which in many cases followed the unimodal approach by concentrating either on visual information either in form of pictures or numbers to identify defects and judge the quality [1], [2], [3]. An obvious example is a deep learning system that employs images. The learning models for instance, Convolutional Neural Networks (CNNs) have been used in milk classification depending on the visual features and dispersion patterns, such as color changes, curdling, and sedimentation [1], [2]. Although useful for clear signs of spoilage, their performance is often disappointing when the changes are subtle visually or intermediate [1]. On the other hand, numerical data such as pH, temperature, storage time, and odor intensity have been used in traditional classification models like Random Forests and Support Vector Machines to predict spoilage levels [4]. These methods are employed in case of detection of physicochemical changes in milk, but they do not possess the capability of judging visual degradation, which is often the first sign noticed by milk consumers.

To address these gaps, researchers have recently explored the use of multimodal deep learning whereby data from multiple source modality is exploited to enhance the classification accuracy and robustness. Multimodal systems are especially useful when single-modality data can be inadequate. For instance, in the area of medical decision making, the integration of MRI images with clinical or genetic information is reported to lead to enhanced early disease diagnosis [5], [6], [7]. The same logic applies in the food industry where researchers have investigated multimodal deep learning for other food-related contexts and has been found to be successful, especially when merging sensor-based data with computer vision methods. This gives capabilities for individual modalities while providing a greater reliability in classification tasks. Research has shown that combination of hyperspectral imaging and temperature increased the accuracy of spoiled fruit detection than individual modality-based inspection [8], [9]. The same multimodal approaches have been investigated for classifying seafood and fruits, combining temperature logs, moisture levels, and spectral data with visual data to strengthen the models [10], [11]. These findings emphasize that multimodal integration improves classification performance and enhances the ability to detect spoilage sooner, especially when a particular type of data is unclear or insufficient on its own.

These developments strengthen the idea that multimodal deep learning forms more robust and flexible systems, particularly in situations where food spoilage has both visible and hidden indicators. In milk quality studies, however, the use of multimodal approaches is not common. As with much research, there is an examination of visual data and numerical datasets independently without blending the two to achieve optimal results. Milk spoilage often occurs in unpredictable ways, early indicators may be chemical, such as pH or the duration of storage, and changes visually occur much later. Therefore, relying solely on visual information can yield wrong conclusions during the early stages of spoilage. This scenario demonstrates the importance of more sophisticated multimodal techniques that integrate what is seen alongside measurement. Moreover, other techniques such as accuracy, F1-score, precision, recall and even confusion metrics dominate most discussions on model performance evaluation which are tightly technical. While these are important for testing how well models work, they don’t reflect how people actually use or understand them. Few studies include human feedback or compare how people make decisions with different types of information. Studying human responses in both unimodal and multimodal setups can provide useful insights into how combining data types affects judgment and decision-making.

# methodology

In this paper, the proposed method will involve creating an online questionnaire using Google Form and an intermediate fusion multimodal deep learning for milk quality analysis via visual and numerical data.

## Google Forms Online Questionnaire

The online questionnaire is organised in three major sections, two of which were designed to cover two distinct types of data modality. The first part collects general information from respondents such as gender, age, and occupation. This information serves to guarantee a broad-based inclusion of participants and to provide the opportunity to examine whether certain social groups exhibit more variety in judgements about milk quality or if they are more biased. For instance, a person’s perception of the taste of food might vary according to their age or professional experience. Respondents are also asked about their milk consumption frequency which is categorized as ‘Rarely’, ‘Average’, or ‘Frequent’. Those who consume milk more regularly may be more accustomed to detecting subtle signs of spoilage, making this question valuable in assessing familiarity with milk quality. Collecting this general information enables a more nuanced interpretation of the results, particularly when comparing human judgment across varying data types.

In the second part of the survey, participants are presented with a selection of milk samples, each examplified through photographs of milk in a clear cup. This portion strived to assess the respondents’ competence to pinpoint degrees of deterioration solely based on visible cues. A sum of nine examples is given, and respondents are guided to categorize each one as either ‘Good’, ‘Spoiling’, or ‘Spoiled’ without having access to any supplementary data from other modalities. A representative of the image employed in the questionnaire is depicted in Figure 1(a) underneath.

In the third section, milk samples with matching visual and numerical information are presented to the participants. The included numerical predictors are sample time, pH value, temperature, storage condition, exposure, and odor intensity. This section is based on the multimodal model, giving a global view on the sample’s condition to support the classification step. The objective is to test the respondents' capacity of rating the quality of milk using both image and numerical information. As in the previous section, respondents are asked to categorize each sample as ‘Good’, ‘Spoiling’, or ‘Spoiled’. A total of nine samples are presented, and one example of the combined visual and numerical data is illustrated in Figure 1(b) below.

|  |  |
| --- | --- |
| (a) | (b) |

**Figure 1.** Example of questions for milk qualitative analysis via Google Forms

## Dataset Preparation

Both visual and numerical datasets are self-collected and obtained simultaneously during the same sampling session to ensure that both data accurately represent the same milk condition. This method was used to keep the data consistent across both modalities and to avoid any mismatches that might happen if the data came from different times or sources. For visual dataset, top-view images of milk in cartons and glass cups are captured using a high-resolution camera while numerical features such as sample time, pH value, temperature, storage condition, exposure status, and odor are recorded simultaneously for numerical dataset. In addition to the milk quality classes which are ‘Good’, ‘Spoiling’, and ‘Spoiled’, this study introduces an additional class called ‘Others’, which consists of images and information unrelated to milk. Since image and numerical features need to align for intermediate fusion, each image is given a sample id that matches each numerical data. Table 1 shows the dataset size for both visual and numerical dataset after splitted into training, validation and testing.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TABLE 1.** Dataset size for visual and numerical dataset | | | | |
|  | **Good** | **Spoiling** | **Spoiled** | **Others** |
| **Training (88%)** | 5,646 | 1,485 | 4,233 | 1,070 |
| **Validation (6%)** | 380 | 97 | 343 | 86 |
| **Testing (6%)** | 397 | 86 | 355 | 50 |

## Developing Intermediate Fusion Multimodal Model

A diagram of different colored squares

AI-generated content may be incorrect.The proposed custom intermediate fusion layer is shown in Figure 2. Key features from both visual and numerical data are extracted and combined before being passed into a custom training layer. For visual data, a custom Convolutional Neural Network (CNN) is used to extract important image features like texture, edges, and patterns. The custom CNN layer is composed of several convolutional layers and subsequent max pooling layers.

**Figure 2.** Custom intermediate fusion layers

CNN is used to capture the high-level spatial features which are important in relation to milk quality. In particular, the image input is fed through a convolutional layer with 32 filters, followed by a max pooling layer, which helps reduce spatial dimensions while retaining important feature information. Hierarchical feature extraction is also proved by the second convolutional layer employing 64 filters followed by another max pooling layer. For numerical data, a dense neural network is used. In a fully connected dense layer, 128 neurons that have been activated by ReLU process the numerical input, which includes features such as sample time, pH, temperature, storage conditions, exposure status, and odor. This layer captures the important statistical features and trends of the numerical dataset. The feature vectors from both CNN and dense network are merged using an intermediate fusion method to form a single combined representation. The combined feature set is then passed through a fully connected neural network, which first activates a dense layer of 256 neurons using ReLU. To avoid overfitting, a dropout layer with a dropout rate of 0.4 is then added. This value was chosen because it helps the model generalize better by randomly turning off some neurons during training without losing too much important information. The final classification is performed using a softmax layer, which outputs probabilities for each of the four milk quality classes.

# results

## Google Forms Online Questionnaire Feedback

In the milk qualitative analysis, there are a total of 93 respondents who participated in the online survey. The demographic information for the respondents is shown in Table 2 below while Table 3 presents a comparative analysis between the ground truth labels and human classification accuracy.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **TABLE 2.** Respondents’ demographic information | | | | | |  |
| **Gender** | | | **Age** | | | | |
| **Male** | | **Female** | **7 - 12** | **13 - 17** | **18 - 25** | **26 - 40** | **Above 40** |
| 30 | | 63 | 5 | 6 | 50 | 24 | 8 |

As shown in Table 3 above, adding numerical data like sample time, pH, temperature, storage, exposure, and odor noticeably improves how accurately people classify milk quality. For example, in Question 3, accuracy jumped from 31.2% with just an image to 76.3% when numerical data was added, an increase of 45.1%. Similar improvements were seen in Questions 1, 5, and 8, with accuracy increases of 27.9%, 30.1%, and 35.5% respectively, showing that numerical data help people make better decisions. One exception was Question 7, where accuracy dropped from 92.5% to 73.1% when numerical data was added. This suggests that in some cases, extra data might confuse respondents, especially when the image alone is already clear. Still, overall, the results support the idea that combining visual and numerical data helps people make more accurate judgments about milk quality.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TABLE 3.** Ground truth VS human analysis | | | | |
| **Question No.** | **GROUND TRUTH** | **Accuracy (%)** | | **Increase in accuracy (%)** |
| **Image only** | **Image & Numerical** |
| 1 | GOOD | 66.7 | 94.6 | 27.9 |
| 2 | SPOILING | 33.3 | 43 | 9.7 |
| 3 | SPOILED | 31.2 | 76.3 | 45.1 |
| 4 | SPOILING | 35.5 | 55.9 | 20.4 |
| 5 | GOOD | 48.4 | 78.5 | 30.1 |
| 6 | SPOILED | 72 | 93.5 | 21.5 |
| 7 | SPOILED | 92.5 | 73.1 | -19.4 |
| 8 | GOOD | 33.3 | 68.8 | 35.5 |
| 9 | SPOILING | 23.7 | 55.9 | 32.2 |

## Intermediate Fusion Multimodal Model

In intermediate fusion, a single model with the proposed layers is trained using different epoch and batch size settings. Table 4 below shows the training, validation, and testing accuracy for each combination. Based on testing accuracy, the best performance for the intermediate fusion model was achieved with 25 epochs and a batch size of 128, reaching 98.87%. This suggests that this combination offers a good balance between learning effectively and avoiding overfitting. In other words, the model performs well not just during training but also when tested on new data. A stable model keeps high performance during training and evaluation, while a model that generalizes well can make accurate predictions on unseen data. As the number of epochs increases, training accuracy improves slightly, showing the model continues to learn. However, more epochs don’t always improve validation and testing accuracy. For example, at 50 epochs, testing accuracy dropped to 98.20%, suggesting overfitting, where the model learns noise instead of useful patterns. In terms of batch size, 128 consistently gives better testing accuracy across all epoch values, especially at 25 epochs. Smaller batch sizes like 32 still perform well, but not as consistent. Although batch size 128 maintains strong training results, small changes in validation accuracy show that batch size can influence how well the model generalizes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **TABLE 4.** Intermediate fusion results | | | | | |
| **Epoch** | **Batch Size** | **Training Accuracy (%)** | | **Validation Accuracy (%)** | **Testing Accuracy (%)** |
| 10 | 32 | 99.79 | 98.90 | | 98.54 |
| 64 | 99.88 | 99.34 | | 98.65 |
| 128 | 99.69 | 99.34 | | 98.76 |
| 25 | 32 | 99.76 | 98.79 | | 98.54 |
| 64 | 99.97 | 99.56 | | 97.41 |
| 128 | 99.99 | 99.34 | | 98.87 |
| 50 | 32 | 99.98 | 98.79 | | 98.20 |
| 64 | 99.96 | 98.57 | | 97.97 |
| 128 | 99.97 | 98.90 | | 98.20 |

# conclusion

In conclusion, this study shows that using multimodal deep learning by combining both visual and numerical data, can improve the accuracy of milk quality classification. By merging features from images and numerical values like sample time, pH level, temperature, storage conditions, exposure status, and odor, we built a model that performs much better than using either type of data alone. The best result came from the combination of 25 epochs and 128 batch sizes, which achieved 98.87% accuracy. The findings also match human feedback. In a survey, humans were better at judging milk quality when they had both image and numerical information, showing that multimodal data helps both humans and machines make better decisions. Overall, this study supports the idea that multimodal systems are more reliable for real-world food quality checks. Future research could explore using this approach in real-time systems or combining it with smart devices for even more accurate and practical monitoring. The proposed intermediate model's functionality can also be applied to other perishable foods, like fruits, meat, and seafood, where spoilage detection is just as important. Expanding the scope of this multimodal framework could have significant benefits for both public health and the food sector.

# Acknowledgments

The authors would like to express their sincere gratitude to the Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka (UTeM) and the Faculty of Engineering and Technology, Multimedia University (MMU) for their support and for providing the facilities and resources essential to this research. Special appreciation is also extended to the laboratory staff and academic supervisors whose guidance and assistance were instrumental in the successful completion of this work.

# References

1. A. M. Pérez-Calabuig, S. Pradana-López, S. Lopez-Ortega, K. de J. Beleño Sáenz, J. C. Cancilla, and J. S. Torrecilla, ‘Application of residual neural networks to detect and quantify milk adulterations’, *Journal of Food Composition and Analysis*, vol. 122, Sep. 2023, doi: 10.1016/j.jfca.2023.105427.

2. P. N. F. F. Kamarudin, N. M. Z. Hashim, M. M. Ibrahim, and M. D. Sulistiyo, ‘Milk Spoilage Classification through Integration of RGB and Thermal Data Analysis’, *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 1839–1851, Jan. 2024, doi: 10.12785/ijcds/1501129.

3. R. R. Kumar, S. Mudepalli, N. Sathish, T. Upender, and S. Mohmmad, ‘An Optimized Neural Network Model to Predict Milk Quality’, in *Proceedings of the 14th International Conference on Cloud Computing, Data Science and Engineering, Confluence 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 326–331. doi: 10.1109/Confluence60223.2024.10463373.

4. A. Tolba, N. N. Mostafa, A. W. Mohamed, and K. M. Sallam, ‘Hybrid Deep Learning Approach for Milk Quality Prediction’, *Precision Livestock*, vol. 1, pp. 1–13, Jan. 2024, doi: 10.61356/j.pl.2024.1199.

5. J. R. Teoh, J. Dong, X. Zuo, K. W. Lai, K. Hasikin, and X. Wu, ‘Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications’, 2024, *PeerJ Inc.* doi: 10.7717/PEERJ-CS.2298.

6. J. Schilcher, A. Nilsson, O. Andlid, and A. Eklund, ‘Fusion of electronic health records and radiographic images for a multimodal deep learning prediction model of atypical femur fractures’, *Comput Biol Med*, vol. 168, Jan. 2024, doi: 10.1016/j.compbiomed.2023.107704.

7. Y. Li *et al.*, ‘A review of deep learning-based information fusion techniques for multimodal medical image classification’, *Comput Biol Med*, vol. 177, p. 108635, 2024, doi: https://doi.org/10.1016/j.compbiomed.2024.108635.

8. X. F. Y. C. J. S. Tianzhe Jiao Chaopeng Guo, ‘A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications’, *Computers, Materials & Continua*, vol. 80, no. 1, pp. 1–35, 2024, doi: 10.32604/cmc.2024.053204.

9. B. Wang *et al.*, ‘The Applications of Hyperspectral Imaging Technology for Agricultural Products Quality Analysis: A Review’, 2023, *Taylor and Francis Ltd.* doi: 10.1080/87559129.2021.1929297.

10. J. Wang *et al.*, ‘Fusing hyperspectral imaging and electronic nose data to predict moisture content in Penaeus vannamei during solar drying’, *Front Nutr*, vol. Volume 11-2024, 2024, doi: 10.3389/fnut.2024.1220131.

11. C. A. Garillos-Manliguez and J. Y. Chiang, ‘Multimodal Deep Learning and Visible-Light and Hyperspectral Imaging for Fruit Maturity Estimation’, *Sensors*, vol. 21, no. 4, 2021, doi: 10.3390/s21041288.