Design and Application of a Custom Late Fusion Layer for Image-Numerical Milk Quality Analysis

Puteri Nur Farzanah Faghira Kamarudin[1,2], Nik Mohd Zarifie Hashim[1,2*]

[1]Fakulti Technology dan Kejuruteraan Elektronik dan Computer, University Technical Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

[2]Centre for Telecommunication Research and Innovation (CeTRI), University technical Malaysia Melaka, Hang Tuah Jaya,76100 Durian Tunggal, Melaka, Malaysia

***Corresponding author**

DOI: https://dx.doi.org/10.47772/IJRISS.2025.908000643

Received: 20 August 2025; Accepted: 28 August 2025; Published: 25 September 2025

ABSTRACT

This paper presents a custom late fusion multimodal deep learning technique for milk quality classification by integrating visual and numerical features. Top-performing unimodal models such as MobileNet, Inception V3, and DenseNet for visual data, and LightGBM, CatBoost, and XGBoost for numerical data were identified through comparative evaluation. The proposed concatenation-with-proposed-layers fusion model achieved a peak testing accuracy of 99.77%, matching or surpassing alternative fusion techniques while employing fewer layers for improved computational efficiency. Comparative experiments demonstrated superior performance over max pooling, majority voting, and weighted average methods, with notable robustness across nine visual–numerical model pairings. A human-centered study further validated the approach, showing that combining visual and numerical inputs improved classification accuracy by up to 45.1% in certain cases. The results highlight the proposed model's effectiveness, stability, and applicability in quality control and safety-critical domains, with potential extension to other multimodal classification tasks requiring high precision.

Keywords— Classification, Data fusion, Late fusion technique, Milk quality, Multimodal Deep Learning.

INTRODUCTION

Ensuring the quality and safety of milk is a critical aspect of public health and food industry standards [1]. Traditional methods of milk quality assessment, such as laboratory testing and sensory evaluation, are often time-consuming, costly, and subjective [1], [2], [3]. With the increasing availability of sensor technologies and digital imaging, automated milk quality analysis using artificial intelligence (AI) has gained significant attention [2], [4], [5].

Deep learning, in particular, has shown promise in handling both image and numerical data for classification tasks [6], [7]. However, most existing studies rely on unimodal approaches [1], [8], focusing either on visual cues such as color and texture [5], [9] or numerical features like pH, temperature and storage conditions separately [2]. These single-modality models may overlook important cross-modal correlations, which can limit their classification performance, especially in complex real-world environments.

To address these limitations, multimodal deep learning has emerged as a powerful solution by integrating multiple types of data. Among the various fusion strategies, late fusion offers greater flexibility by allowing each modality to be processed independently before being combined [10], [11]. Despite its advantages, most late fusion implementations in current literature are typically employed as fixed, non-trainable operations that do not adapt to the characteristics of the input modalities or the specific demands of the task. To address this limitation, we propose a custom late fusion layer that embeds these operations within a configurable architecture. This approach enables the model to perform modality integration in a more adaptive and learnable manner, thereby enhancing its ability to extract and combine complementary features from both visual and numerical inputs for improved milk quality classification.

The contributions of this study are as follows:

1. A comparative analysis of unimodal models for visual and numerical milk quality data.
2. The design of a custom late fusion layer for multimodal integration.
3. An evaluation showing excellent classification performance and alignment with human judgment, validating the effectiveness of the proposed method.

Dataset Acquisition and Preparation

The preparation of these samples is conducted in a controlled environment, to reduce the influence of external factors such as contamination, changes in temperature, or inconsistent lighting, which could affect the quality and consistency of the collected data. Both the visual and numerical datasets were self-collected concurrently during the same sampling sessions to ensure that each modality accurately reflects the same milk condition. This approach was adopted to maintain consistency between modalities and to prevent discrepancies that could arise from data collected at different times or from separate sources.

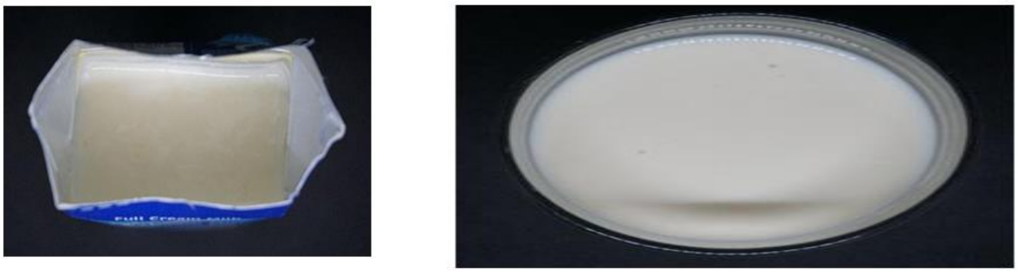

Fig. 1 Top-view image of milk: (a) in carton and (b) in glass cup

The visual dataset comprises top-view images of milk contained in cartons and glass cups as shown in Fig. 1(a) and (b) respectively. They were captured using a Sony α6000 E-mount camera with APS-C Sensor along with an SELP1650 interchangeable lens. Simultaneously, the numerical dataset includes recorded features such as sample time, pH level, temperature, storage condition, exposure status, and odor. In addition to the primary milk quality classes which are ‘Good’, ‘Spoiling’, and ‘Spoiled’, an extra class labeled ‘Others’ was introduced to include images and data unrelated to milk. For effective intermediate fusion, each image is assigned a sample ID that corresponds to its respective numerical data, ensuring proper alignment between the modalities.

Table I provides a sample of the recorded numerical data, while Table II presents the coding scheme used to represent storage condition, exposure status, and odor in numerical form. Table III presents the dataset distribution across training, validation, and testing sets for both visual and numerical inputs.

TABLE I SAMPLE OF THE RECORDED NUMERICAL DATA

Sample Time	24	48	72	96
pH	5.62	6.29	6.38	-999
Temp. (°C)	30.25	18.23	31.51	-999
Storage	0	1	0	-999
Exposure	1	0	2	-999
Odor	1	0	1	-999
Class	Spoiled	Good	Spoiling	Others

TABLE II NUMERICAL REPRESENTATION FOR STORAGE CONDITION, EXPOSURE STATUS, AND ODOR

Numerical Representation	Storage	Exposure	Odor
0	Room temperature	Perfectly sealed	No odor
1	Refrigerator	Breached seal	Slight odor
2	-	Fully opened	Strong odor

TABLE III DATASET DISTRIBUTION 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

Visual Analysis

In this study, a range of well-established deep learning architectures are employed as baseline models for visual classification. These include AlexNet, DenseNet, Inception V1, Inception V3, LeNet-5, MobileNet, ResNet, VGG16, and VGG19. The model that demonstrates the highest classification accuracy will be selected as the final architecture for integration within the multimodal deep learning framework.

Pre-trained versions of each model are utilized to perform milk quality analysis using the visual dataset. Using pre-trained models enables the use of feature representations learned from large-scale image datasets, which enhances classification accuracy and reduces training time, particularly beneficial when dealing with limited data. To ensure a fair and unbiased comparison among models, all hyperparameters are held constant. The epoch, optimizer and learning rates for all models were set to 10, Adam optimizer and 0.0001 respectively. This uniformity eliminates discrepancies caused by individual model tuning and ensures that any differences in performance are attributed solely to the model architecture rather than varying experimental conditions.

Referring to Fig. 2, among the evaluated models, MobileNet achieved the highest classification accuracy at 99.66%, followed by Inception V3 at 99.44% and DenseNet at 99.1%. These architectures outperformed others in effectively capturing critical visual features relevant to milk quality classification. In contrast, ResNet recorded the lowest accuracy at 52.93%, indicating its limited suitability for this dataset. Based on these findings, MobileNet, Inception V3, and DenseNet were selected for integration into the late fusion multimodal framework. Their strong performance and ability to generalize key visual cues make them well-suited for contributing to a robust and accurate multimodal classification system.

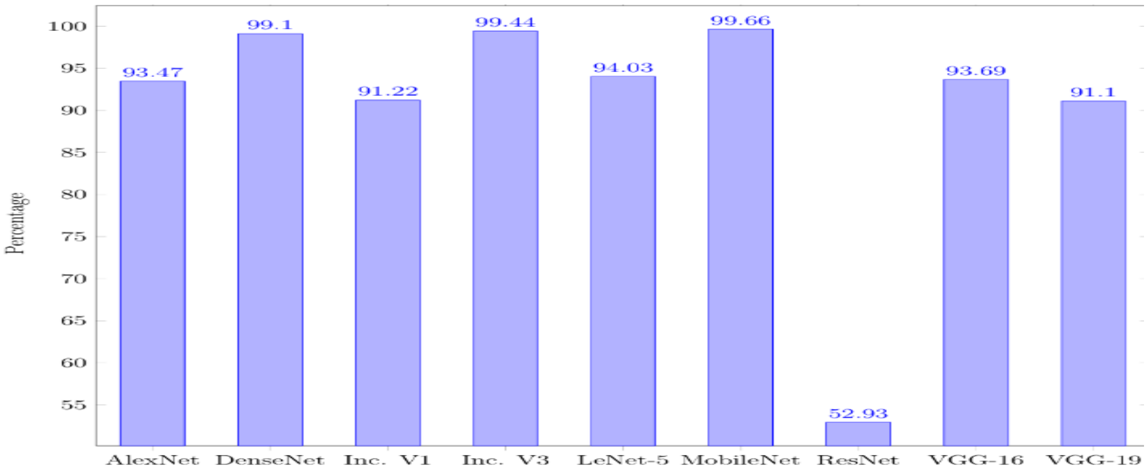

Fig. 2 Visual models' accuracy

Numerical Analysis

For numerical data analysis, eight baseline models are selected which are CatBoost, K-Nearest Neighbors (K-NN), LightGBM, Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine (SVM), and XGBoost. These models are chosen due to their proven effectiveness in a wide range of classification applications. Similar to visual analysis, the model that demonstrates the highest classification accuracy will be selected as the final architecture for integration within the multimodal deep learning framework. By assessing the accuracy of these models, the most suitable one for the current classification task can be determined. Pre-trained versions of each model were employed to classify milk quality based on the numerical dataset.

Among the models evaluated as shown in Fig. 3 below, LightGBM achieved the highest accuracy at 96.85%, closely followed by CatBoost at 96.51%. Both Random Forest (RF) and XGBoost recorded identical accuracies of 96.28%. These models outperformed the others due to their inherent capability to effectively process structured, tabular data. Gradient boosting algorithms such as LightGBM, CatBoost, and XGBoost are particularly well-suited for capturing complex, non-linear relationships and interactions among numerical features such as sample time, pH, temperature, storage condition, exposure status, and odor which are critical indicators in assessing milk quality.

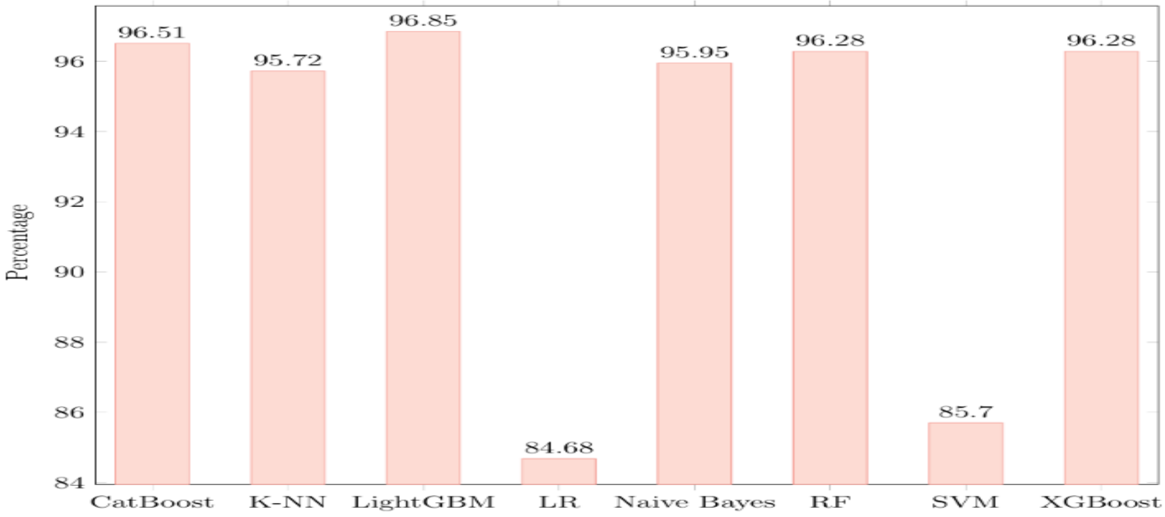

Fig. 3 Numerical models’ accuracy

As the two highest-performing models in this analysis, LightGBM and CatBoost are directly selected for inclusion in the final late fusion model. For the third model, Random Forest and XGBoost demonstrated identical testing accuracies, prompting a comparison of their training and validation performance to determine the more suitable option. Table IV presents a comparison of their training and validation accuracies.

TABLE IVTRAINING AND VALIDATION ACCURACY FOR RANDOM FOREST AND XGBOOST

	Random Forest	XGBoost
Training Accuracy (%)	99.84	100
Validation Accuracy (%)	96.36	96.25

Based on the above analyses, LightGBM, CatBoost, and XGBoost are identified as the top three performing numerical models and are therefore selected for the development of the late fusion model.

Late Fusion Multimodal Model Development

Late fusion, also referred to as decision fusion, involves combining the outputs of models trained on different data modalities [10], [12], [13]. This approach uses the strengths of each individual model, making it particularly effective when models excel in handling specific types of input data [12], [13]. A common use case of late fusion can be found in sentiment analysis, where textual, audio, and visual information are independently processed and their results integrated to produce a final sentiment classification [14], [15]. In this paper, a custom concatenation-based late fusion method is proposed and implemented through specially designed layers. This approach is evaluated against other late fusion techniques, including standard concatenation, max pooling, majority voting, and weighted averaging.

Following the independent training of nine visual models and eight numerical models, the top three highest-performing models from each category (visual and numerical analysis) are selected for late fusion multimodal analysis. Selecting the top three provides greater flexibility in exploring various fusion combinations while maintaining a focus on the most promising candidates. This strategy also minimizes the risk of depending solely on a single model that may not perform consistently across different scenarios. The selected models are then systematically paired to assess their performance using the proposed concatenation-based late fusion technique. In total, nine unique model pairings as shown in Table V are evaluated to determine which combination achieves the highest classification accuracy.

TABLE V LATE FUSION UNIQUE MODEL PAIRING (TOP THREE FROM VISUAL AND NUMERICAL ANALYSIS)

Model Pairing	
Visual	Numerical
MobileNet	LightGBM
MobileNet	CatBoost
MobileNet	XGBoost
DenseNet	LightGBM
DenseNet	CatBoost
DenseNet	XGBoost
Incp. V3	LightGBM
Incp. V3	CatBoost
Incp. V3	XGBoost

Once the visual model is trained from scratch using the visual dataset and the numerical model is trained using the numerical dataset, their respective feature representations are extracted and fed into the late fusion model. The visual dataset is processed through the trained visual model, which serves as a feature extractor to capture high-level representations of the images. Likewise, the numerical model generates feature representations based on the output probability distributions, effectively summarizing the key patterns in the numerical data. These extracted features are then concatenated into a unified feature vector, allowing the fusion model to make use of the complementary information from both modalities for excellent classification performance.

Concatenation with Proposed Layers

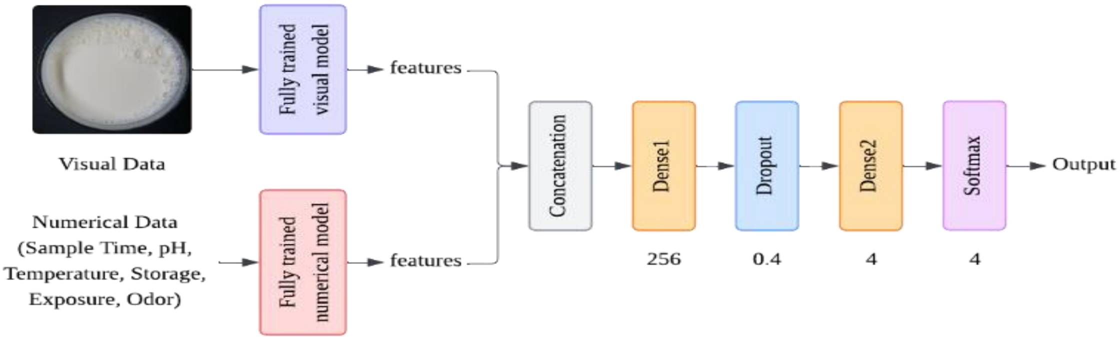
Visual Data
Fully trained visual model
features
Numerical Data (Sample Time, pH, Temperature, Storage, Exposure, Odor)
Fully trained numerical model
features
Concatenation
Dense1
256
Dropout
0.4
Dense2
4
Softmax
4
Output

Fig. 4 Concatenation with proposed layers

In the first late fusion model as shown in Fig. 4, which utilizes concatenation with proposed layers, the concatenated features are passed through a fully connected layer consisting of 256 neurons with a ReLU activation function. This is followed by a dropout layer with a dropout rate of 0.4 to reduce the risk of overfitting. Finally, a softmax output layer with four neurons is applied to produce probability distributions across the four target classes.

Standard Concatenation

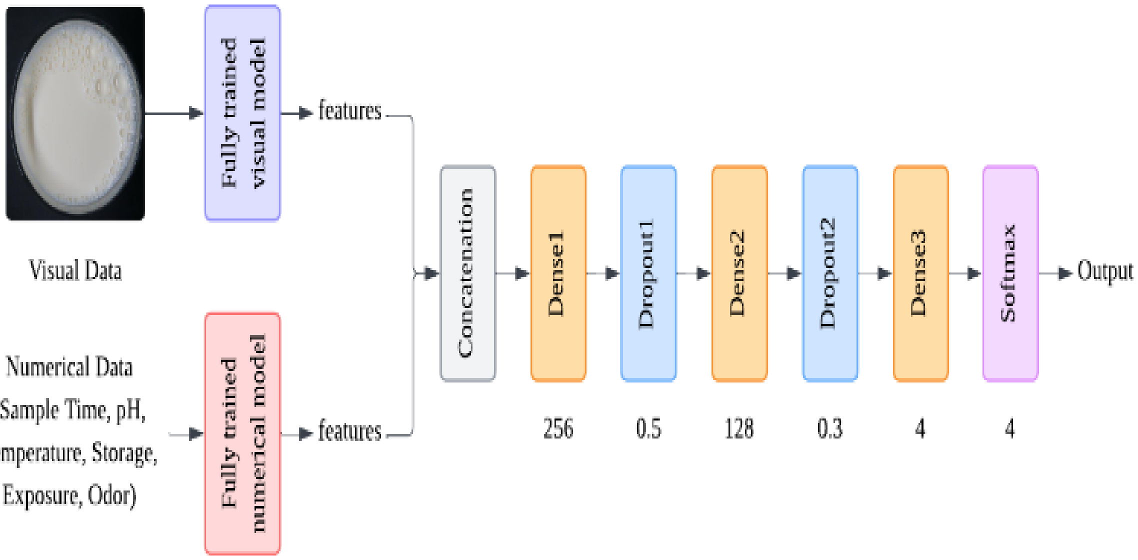
Visual Data
Fully trained visual model
features
Numerical Data (Sample Time, pH, Temperature, Storage, Exposure, Odor)
Fully trained numerical model
features
Concatenation
Dense1
Dropout1
Dense2
Dropout2
Dense3
Softmax
Output
256
0.5
128
0.3
4
4

Fig. 5 Standard concatenation

The second late fusion model, which uses a standard concatenation approach, is illustrated in Fig. 5. In this model, the concatenated features are processed through a series of fully connected layers. It begins with a dense layer of 256 neurons activated by a ReLU function, followed by a dropout layer with a rate of 0.5 to mitigate overfitting. This is followed by another dense layer with 128 neurons and an additional dropout layer with a 0.3 dropout rate. Finally, a dense layer with four neurons is applied, and the output is passed through a softmax activation function to produce probability distributions across the four target classes.

Max Pooling with Proposed Layers

The third late fusion model follows the same architecture as the first model which is concatenation with proposed layers, but it includes an additional global max pooling layer applied to the concatenated features. This modified structure is illustrated in Fig. 6 below.

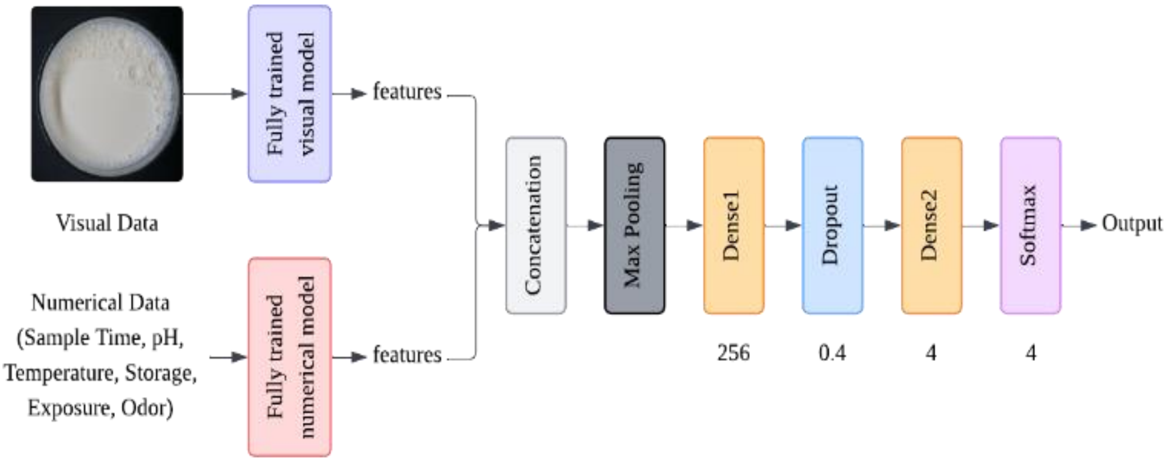
Visual Data
Fully trained visual model
features
Numerical Data (Sample Time, pH, Temperature, Storage, Exposure, Odor)
Fully trained numerical model
features
Concatenation
Max Pooling
Dense1
Dropout
Dense2
Softmax
Output
256
0.4
4
4

Fig. 6 Max pooling with proposed layers

Majority Voting

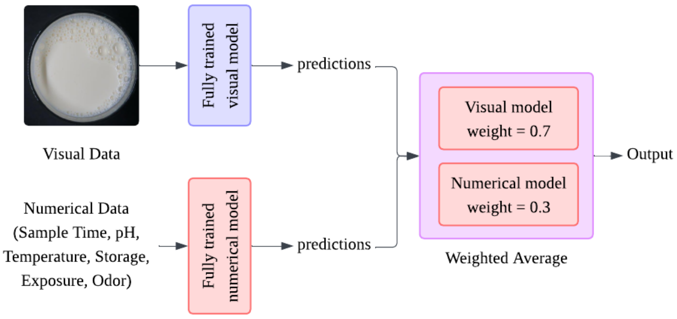
Visual Data
Fully trained visual model
predictions
Numerical Data (Sample Time, pH, Temperature, Storage, Exposure, Odor)
Fully trained numerical model
predictions
Visual model weight = 0.7
Numerical model weight = 0.3
Weighted Average
Output

Fig. 7 Majority Voting

The fourth late fusion models differ from the previous models by utilizing the model's final predictions instead of their extracted features. As illustrated in Fig. 7, the majority voting model receives output from two independently trained models, each processing a different modality. Instead of fusing feature-level representations, this approach combines the final predictions through a soft voting mechanism, aggregating the predicted class probabilities to determine the overall classification result.

Weighted Average

Fig. 8 below depicts the weighted average fusion technique, which, similar to majority voting, combines predictions from two independently trained models, one for each modality. In this approach, the visual model is given a higher weight of 0.7, while the numerical model is assigned a weight of 0.3. These weights were selected empirically, with greater emphasis placed on the modality expected to contribute more informative features to the classification task. The visual modality is prioritized due to its ability to capture rich physical characteristics of the milk samples, whereas the numerical modality, although structured, provides a more limited feature set. This weighting strategy ensures that the fusion process makes use of the strengths of each modality to enhance the final prediction accuracy.

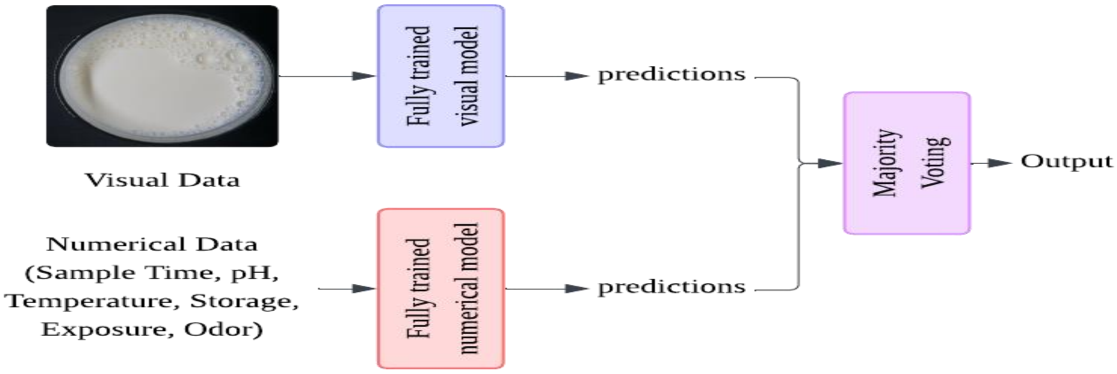
Visual Data
Fully trained visual model
predictions
Numerical Data (Sample Time, pH, Temperature, Storage, Exposure, Odor)
Fully trained numerical model
predictions
Majority Voting
Output

Fig. 8 Weighted Average

RESULTS AND DISCUSSION

This section will present the results of late fusion multimodal specifically concatenation technique with proposed custom layers, standard concatenation, max pooling, majority voting, and weighted average. Based on the results presented in visual and numerical analysis, MobileNet achieved the highest accuracy among the visual classification models, while LightGBM outperformed other models in numerical classification. Hence, both models were used for the analysis of late fusion multimodal model.

Concatenation with Proposed Layers

Table VI below displays the accuracy percentages across various epoch and batch size combinations for the first model which is concatenation with proposed csutom layers. The highlighted row in Table 6 indicates the specific combination that yielded the highest classification accuracy.

TABLE VI MobileNet X LightGBM (Concatenation with Proposed Layers)

Epoch	Batch Size	Training Acc. (%)	Validation Acc. (%)	Testing Acc. (%)
10	32	99.99	100	99.77
	64	100	100	99.77
	128	99.99	100	99.66
25	32	100	100	99.77
	64	100	100	99.77
	128	100	100	99.77
50	32	100	100	99.77
	64	100	100	99.77
	128	100	100	99.77

The optimal combination for this model was identified as 10 epochs with a batch size of 64, achieving 100% training and validation accuracy and 99.77% testing accuracy. This setting offers an effective balance between generalization and stability. Increasing the number of epochs did not yield notable accuracy improvements, as training, validation, and testing accuracy consistently remained around 100% and 99.77%, respectively. This indicates that the model can effectively learn dataset patterns within a relatively small number of epochs, and higher epoch counts neither enhance nor degrade performance. Similarly, variations in batch size across all epochs had no measurable impact on testing accuracy, demonstrating that the concatenation-with-proposed-layers model maintains strong generalization regardless of batch size.

To assess the effectiveness of the proposed custom late fusion model, the rest of model pairings as listed in Table 5 were analysed. This design choice ensures that performance differences across the nine model pairs are solely attributable to the choice of base models, rather than variations in the fusion strategy. By applying the same fusion mechanism to all pairings, the results directly reflect the robustness and adaptability of the proposed custom model across diverse combinations of visual and numerical classifiers. As the optimal combination of epoch and batch size of MobileNet X LightGBM are 10 and 64 respectively, hence, the combination is used for the rest of model pairings. The model pairings' results are as tabulated in Table VII below.

TABLE VII Other model pairing at epoch=10, batch size=64 (Concatenation with Proposed Layers)

Model Pairing	Training Acc. (%)	Validation Acc. (%)	Testing Acc. (%)
MobileNet X CatBoost	99.99	100	99.77
MobileNet X XGBoost	99.99	100	99.77
DenseNet X LightGBM	99.85	99.89	99.32
DenseNet X CatBoost	99.96	100	99.77
DenseNet X XGBoost	99.85	99.89	99.44
Incp. V3 X LightGBM	99.91	99.89	99.66
Incp. V3 X CatBoost	99.98	99.89	99.77
Incp. V3 X XGBoost	99.90	99.89	99.77

Across all model pairings, training and validation accuracy consistently approached 100%, while testing accuracy was slightly lower, ranging from 99.32% to 99.77%. The highest testing accuracy of 99.77% was achieved by multiple pairings, including MobileNet with CatBoost and XGBoost, DenseNet with CatBoost, and Inception V3 with CatBoost and XGBoost matching with MobileNet X LightGBM. This consistency indicates that the concatenation-with-proposed-layers model effectively integrates complementary features from visual and numerical modalities. Moreover, MobileNet X LightGBM outperformed certain LightGBM-based combinations, such as DenseNet X LightGBM and Inception V3 X LightGBM, which achieved slightly lower accuracies of 99.32% and 99.66% respectively.

Other Late Fusion Techniques

Table VIII below displays the accuracy percentages for other late fusion techniques which are standard concatenation, max pooling, majority voting, and weighted average.

TABLE VIII OTHER MODEL PAIRING AT EPOCH=10, BATCH SIZE=64 (CONCATENATION WITH PROPOSED LAYERS)

Model Pairing	Training Acc. (%)	Validation Acc. (%)	Testing Acc. (%)
Standard Concatenation	100	100	99.77
Max Pooling	45.41	41.94	44.71
Majority Voting	N/A	N/A	99.77
Weighted Average	N/A	N/A	82.55

A comparison between the concatenation-with-proposed-layers model and the standard concatenation model shows identical accuracies for training, validation, and testing which are 100%, 100%, and 99.77%, respectively. Notably, the proposed model employs a reduced number of layers. This indicates that comparable performance can be achieved without additional architectural complexity. By eliminating unnecessary layers, the proposed model offers a more computationally efficient solution, reducing training time and memory requirements while maintaining state-of-the-art accuracy.

In contrast, the max pooling model achieves only 45.41% training accuracy, 41.94% validation accuracy, and 44.71% testing accuracy. This substantial performance gap indicates that the concatenation-with-proposed-layers model offers markedly superior generalization and learning capabilities.

Although both the concatenation-with-proposed-layers model and the voting ensemble achieve the same testing accuracy of 99.77%, training and validation accuracy data are unavailable for the voting ensemble, making it challenging to evaluate its learning behavior and generalization ability. The voting ensemble, however, is exceptionally fast, with an approximate training time of 0 seconds. This indicates that while the voting ensemble offers high computational efficiency, its performance during the training and validation phases remains unclear. If the aim is to maximize precision with full transparency across training, validation, and testing, the concatenation-with-proposed-layers method is the preferable choice. Conversely, the voting ensemble is a strong alternative when rapid prediction is the priority, provided that the model's generalization capability has already been established.

Similar to majority voting, the weighted average fusion method demonstrates efficient training time, making it suitable for real-time applications where quick results are essential. However, this speed advantage comes at the cost of precision. Its testing accuracy of 82.55% is notably lower than the 99.77% achieved by the concatenation-with-proposed-layers model. Furthermore, the absence of training and validation accuracy data makes it challenging to assess the model's generalization capability or consistency in learning.

Overall, while the weighted average approach offers simplicity and speed, it underperforms in prediction accuracy. It may be practical in scenarios where speed outweighs precision or when computational resources are limited. In contrast, the concatenation-with-proposed-layers model delivers far greater robustness and reliability, making it the preferred choice for applications that demand high classification accuracy, such as quality control or safety-critical systems.

Human Analysis for Unimodal VS Multimodal

This paper also provides a qualitative validation of multimodal deep learning through a human-centered experiment conducted using Google Forms. The survey was divided into three parts. The first part gathered general demographic details from respondents, including gender, age, occupation, and frequency of milk consumption. In the second part, participants were asked to classify nine milk samples based solely on visual information. The third part presented the same samples along with both visual and numerical data, such as sampling time, pH level, temperature, storage condition, exposure status, and odor. For each sample, respondents are required to classifiy it as either 'Good', 'Spoiling', or 'Spoiled'. Table IX below presents a comparative analysis between the ground truth labels and human classification accuracy.

TABLE IX GROUND TRUTH VS HUMAN ANALYSIS

No.	GROUND TRUTH	Accuracy (%)		Increase in accuracy (%)
		Image only	Image & Numerical	
1	Good	66.7	94.6	1
2	Spoiling	33.3	43	2
3	Spoiled	31.2	76.3	3
4	Spoiling	35.5	55.9	4
5	Good	48.4	78.5	5
6	Spoiled	72	93.5	6
7	Spoiled	92.5	73.1	7
8	Good	33.3	68.8	8
9	Spoiling	23.7	55.9	9

As presented in Table IX, incorporating numerical data such as sample time, pH, temperature, storage condition, exposure status, and odor significantly enhances the accuracy of human milk quality classification. For instance, in Question 3, accuracy rose from 31.2% using only visual data to 76.3% when numerical information was included, marking a 45.1% improvement. Comparable gains were observed in Questions 1, 5, and 8, with increases of 27.9%, 30.1%, and 35.5% respectively, indicating that numerical features support better decision making. However, an exception occurred in Question 7, where accuracy declined from 92.5% to 73.1% upon adding numerical data. This suggests that in certain cases, supplementary information may cause confusion, particularly when visual cues alone provide a clear assessment. Nevertheless, the overall findings reinforce the conclusion that combining visual and numerical inputs improves judgment accuracy in milk quality evaluation.

CONCLUSION

This paper demonstrated the effectiveness of a multimodal deep learning model for milk quality classification by integrating visual and numerical features through a custom late fusion technique. For visual classification, MobileNet, Inception V3, and DenseNet emerged as the top-performing architectures, achieving accuracies of 99.66%, 99.44%, and 99.10% respectively, with MobileNet selected as the primary visual model in the fusion framework. In the numerical analysis, LightGBM, CatBoost, and XGBoost achieved the highest accuracies, all exceeding 96%, confirming their suitability for capturing complex relationships in structured data such as pH, temperature, and odor.

The proposed concatenation-with-proposed-layers fusion model proved highly effective, consistently delivering near-perfect training and validation accuracy and achieving a peak testing accuracy of 99.77% across multiple visual–numerical pairings. Comparative analysis with alternative fusion techniques showed that max pooling and weighted average fusion significantly underperformed in prediction accuracy, while majority voting achieved high testing accuracy but lacked transparency in training and validation performance. These findings indicate that the proposed model offers the best balance of precision, robustness, and interpretability, making it well-suited for quality control and safety-critical applications.

Human-centered evaluation further reinforced the advantages of multimodal inputs. The addition of numerical data alongside visual data substantially improved respondents' classification accuracy in most cases, with gains of up to 45.1% for specific samples. While a minor decrease in accuracy occurred in one scenario due to potential information overload, the overall results confirmed that multimodal data presentation enhances decision-making quality.

In conclusion, the integration of high-performing unimodal models into a late fusion architecture with the proposed concatenation method provides a powerful and reliable solution for milk quality classification. The approach demonstrates strong generalization capabilities, high predictive accuracy, and meaningful performance improvements over simpler fusion strategies. Beyond the milk quality field, these findings highlight the potential of multimodal deep learning frameworks to improve classification accuracy and decision support in other applications where both visual and numerical features are critical.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the Faculty Technology dan Kejuruteraan Elektronik dan Computer, University technical Malaysia Melaka (UTeM), for providing the facilities and resources necessary for this research. Sincere thanks are also extended to the laboratory staff and academic supervisors for their invaluable guidance and assistance, which greatly contributed to the successful completion of this work.

REFERENCES

1. W. Zhang, E. Chen, M. Anderson, S. Thompson, and J. Lee, “Multimodal Deep Learning-Based Intelligent Food Safety Detection and Traceability System,” International Journal of Management Science Research, vol. 8, 2024, doi: 10.53469/ijomsr.2025.08(03).09.
2. 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.
3. L. Zhang, Q. Yang, and Z. Zhu, “The Application of Multi-Parameter Multi-Modal Technology Integrating Biological Sensors and Artificial Intelligence in the Rapid Detection of Food Contaminants,” Foods, vol. 13, no. 12, Jun. 2024, doi: 10.3390/foods13121936.
4. E. Buoio, V. Colombo, E. Ighina, and F. Tangorra, “Rapid Classification of Milk Using a Cost-Effective Near Infrared Spectroscopy Device and Variable Cluster–Support Vector Machine (VC-SVM) Hybrid Models,” Foods, vol. 13, no. 20, Oct. 2024, doi: 10.3390/foods13203279.
5. S. Seilov, D. Abildinov, M. Baydeldinov, A. Nurzhaubayev, B. Zhursinbek, and X. G. Yue, “Integration of Electronic Nose and Machine Learning for Monitoring Food Spoilage in Storage Systems,” International Journal of Online and Biomedical Engineering, vol. 20, no. 16, pp. 117–130, Dec. 2024, doi: 10.3991/ijoe.v20i16.52911.
6. M. Mohd Ali, N. Hashim, S. Abd Aziz, and O. Lasekan, “Utilisation of Deep Learning with Multimodal Data Fusion for Determination of Pineapple Quality Using Thermal Imaging,” Agronomy, vol. 13, no. 2, Feb. 2023, doi: 10.3390/agronomy13020401.
7. J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, “Multimodal deep learning models for early detection of Alzheimer's disease stage,” Sci Rep, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-020-74399-w.
8. F.-Z. Nakach, A. Idri, and E. Goceri, “A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification,” Artif Intell Rev, vol. 57, no. 12, p. 327, 2024, doi: 10.1007/s10462-024-10984-z.
9. S. Fan et al., “On line detection of defective apples using computer vision system combined with deep learning methods,” J Food Eng, vol. 286, Dec. 2020, doi: 10.1016/j.jfoodeng.2020.110102.
10. V. A. Torres Caceres, K. Duffaut, A. Yazidi, F. Westad, and Y. B. Johansen, “Automated well log depth matching: Late fusion multimodal deep learning,” Geophys Prospect, vol. 72, no. 1, pp. 155–182, Jan. 2024, doi: 10.1111/1365-2478.13200.

11. 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.
12. T. Jiao, C. Guo, X. Feng, Y. Chen, and J. Song, "A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications," 2024, Tech Science Press. doi: 10.32604/cmc.2024.053204.
13. Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," Multimed Tools Appl, vol. 80, no. 2, pp. 2887–2905, Jan. 2021, doi: 10.1007/s11042-020-08836-3.
14. J. S. Chu and S. Ghanta, "Integrative Sentiment Analysis: Leveraging Audio, Visual, and Textual Data", doi: 10.5121/csit.2024.1402011.
15. I. Khan and S. Kadu, "Sentiment Analysis of Multimodal Content: A Fusion of Visual and Textual Clues," in 2024 International Conference on Innovation and Novelty in Engineering and Technology, INNOVA 2024 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/INNOVA63080.2024.10846981.