Food Freshness Evaluation Using a CLIP-Based Architecture

Md. Siam Ansary, Amina Brinto, and Shaila Sajnin Keya

Abstract—In this work, we present an efficient deep learning framework for automated fresh and stale food classification using transfer learning with a pretrained CLIP-based feature extractor. The proposed system employs frozen vision transformer (ViT) embeddings from CLIP as generalized visual descriptors and integrates them with a lightweight multi-layer perceptron (MLP) classifier for binary classification. To enhance generalization, extensive data augmentation and stratified dataset partitioning were applied to the publicly available Fresh and Stale Classification dataset. Experimental results reveal a consistent improvement across ten training epochs, achieving a final test accuracy of 97.99%, F1-score of 0.9808, and ROC–AUC of 0.9985. The proposed model demonstrates excellent discriminative performance, robust convergence, and strong generalization capabilities while maintaining computational efficiency. These results confirm the suitability of CLIP-based visual representations for high-accuracy food quality assessment and real-time freshness detection applications.

***Keywords*—image classification, clip, food freshness, health.**

NOMENCLATURE

CLIP	Contrastive Language-Image Pre-training.
MLP	Multilayer perceptron.
ViT	Vision Transformers.
CNN	Convolutional Neural Network.
ACC	Accuracy.
P	Precision.
R	Recall.
F1	F1-score.
SP	Specificity.
ROC	Receiver Operating Characteristic.
AUC	Area under the curve.
MCC	Matthews Correlation Coefficient.
CUDA	Compute Unified Device Architecture.
GPU	Graphics Processing Unit.

I. INTRODUCTION

Food quality assessment plays a critical role in ensuring consumer safety, reducing waste, and maintaining supply-chain efficiency. Traditional approaches for freshness detection rely heavily on manual inspection or sensor-based measurements, which are time-consuming, labor-intensive, and prone to human error. Recent advancements in deep learning and computer vision have enabled automated food-quality recognition by leveraging large-scale image datasets and pretrained models. However, many existing solutions depend on fine-tuning deep convolutional networks, which require significant computational resources and domain-specific data.

To address these limitations, this study proposes a transfer-learning–based classification framework that integrates pretrained CLIP (Contrastive Language–Image Pretraining) embeddings with a lightweight Multi-Layer Perceptron (MLP) classifier. CLIP's vision transformer (ViT-B/16) backbone was employed as a fixed feature extractor to capture rich semantic and textural representations from images of food items. The extracted features were subsequently classified using a compact, fully connected MLP network. This approach reduces training cost while maintaining high discriminative performance.

The proposed model was evaluated on the publicly available Fresh and Stale Classification dataset from Kaggle. The data were divided into training, validation, and test subsets using stratified sampling to maintain class balance. The system achieved outstanding accuracy and F1-score, demonstrating the effectiveness of pretrained transformer features for specialized classification tasks. The major contributions of this research are as follows:

1. Development of a CLIP-based feature extraction framework for food freshness classification without end-to-end fine-tuning.
2. Implementation of a lightweight MLP classifier optimized for fast convergence and minimal computational overhead.
3. Comprehensive evaluation using multiple performance metrics including accuracy, precision, recall, F1-score, ROC–AUC, and Matthews correlation coefficient.
4. Empirical validation showing that frozen CLIP embeddings can achieve near state-of-the-art results on a small-scale, domain-specific dataset.

Manuscript received 18 October 2025; revised 30 November 2025.

Md. Siam Ansary is with the Department of Computer Science and Engineering (CSE), Ahsanullah University of Science and Technology, Dhaka, BANGLADESH. (e-mail: siamansary.cse@gmail.com).

Amina Brinto is with the Department of Obstetrics and Gynaecology, Kurmitola General Hospital, Dhaka, Bangladesh. (e-mail: aminabrinto@gmail.com).

Shaila Sajnin Keya is with the Institute of Nutrition and Food Science, University of Dhaka, BANGLADESH. (e-mail: shailasajninkeya1@gmail.com).

Digital Object Identifier (DOI): 10.53907/enpesj.v5i2.344

The experimental findings highlight that transfer learning using CLIP-derived representations provides a powerful and efficient pathway for food-quality classification, opening avenues for deployment in industrial inspection systems, automated retail environments, and mobile consumer applications.

II. Literature Review

Yuan *et al.* [1] worked on vegetable and fruit freshness detection using deep visual features. They extracted CNN-based representations from RGB images and trained a classifier that achieved high recognition accuracy for freshness identification across multiple produce categories.

Gao *et al.* [2] developed a food image classification model utilizing a Vision Transformer (ViT) with extensive data and feature augmentation. Their framework significantly outperformed conventional CNN architectures, showing higher precision on benchmark food datasets.

Jo *et al.* [3] conducted research on fresh meat quality assessment through hyperspectral imaging (HSI). By integrating spatial–spectral CNNs, they predicted physicochemical properties related to meat freshness and demonstrated substantial gains over RGB-only models.

Choi *et al.* [4] proposed a hybrid model combining hyperspectral imaging and chemometric analysis for predicting pork freshness. The model accurately estimated total volatile basic nitrogen (TVB-N) values and sensory freshness scores, confirming the potential of HSI for non-destructive freshness assessment.

Lun *et al.* [5] presented a comprehensive review of deep learning–enhanced spectroscopic technologies for food quality analysis. Their study emphasized the synergy between deep neural networks and spectral sensing methods in evaluating ripeness, adulteration, and spoilage.

Sonwani *et al.* [6] worked on an integrated food spoilage monitoring system employing multiple sensors and machine learning algorithms. The framework successfully detected early spoilage signs under real-world conditions by analyzing volatile gas emissions and environmental parameters.

Shu *et al.* [7] investigated fruit freshness classification using a ResNet-101 backbone enhanced with non-local attention mechanisms. Their model achieved superior recall and F1-score on fruit datasets with varying freshness levels, proving the effectiveness of attention-based CNNs.

Nikzadfar *et al.* [8] reviewed hyperspectral imaging and artificial intelligence integration for food quality and safety. They analyzed recent methods that combine spatial–spectral data with deep learning architectures for rapid and non-invasive freshness detection.

Gatti *et al.* [9] applied CLIP-based transfer learning for visual verification in food packaging. They extracted frozen CLIP embeddings and trained lightweight classifiers to detect mismatches in food order packaging, achieving high accuracy in industrial inspection environments.

Mehdizadeh *et al.* [10] explored AI-driven, non-destructive detection of meat freshness using spectral sensors. Their model correlated deep-learning predictions with chemical indicators of spoilage, reaching ROC–AUC scores above 0.95.

Varga *et al.* [11] investigated fruit ripeness estimation using hyperspectral imaging combined with deep learning. They applied convolutional neural networks to predict ripeness stages with strong generalization across different fruit types.

Anwar *et al.* [12] conducted a review on food quality assessment using machine learning and sensors. Their study concluded that sensor fusion, combining electronic nose, imaging, and spectral data, enhances reliability in freshness classification.

Radford *et al.* [13] introduced CLIP (Contrastive Language–Image Pretraining), which enabled large-scale visual–language representation learning. Their model set a foundation for zero-shot transfer learning, later utilized for various food classification tasks.

Dosovitskiy *et al.* [14] developed the Vision Transformer (ViT) architecture that processes image patches through self-attention mechanisms. ViT demonstrated superior performance on image classification benchmarks and influenced modern food-vision approaches.

Bossard *et al.* [15] proposed the Food-101 dataset and baseline CNN models for food category classification. This dataset has since been widely used for evaluating and fine-tuning deep food recognition systems.

Ghosh *et al.* [16] presented NoisyViT, a robust vision transformer framework for food image recognition under noisy environments. Their approach enhanced classification stability and improved performance on low-quality food imagery.

Liu *et al.* [17] conducted a comprehensive review on deep learning in food image recognition. They discussed CNNs, transformers, and multimodal architectures, highlighting recent advances and challenges in large-scale food datasets.

III. Implementation Methodology

The implementation pipeline was designed to systematically train, validate, and evaluate a deep learning model for binary image classification between *fresh* and *stale* food samples. The overall workflow comprises six major stages: dataset preparation, preprocessing and augmentation, model architecture design, feature extraction, classifier training, and evaluation. Each component was implemented in Python using PyTorch and the `timm` vision library, executed in Google Colab with GPU acceleration.

A. Dataset Preparation

The "Fresh and Stale Classification" dataset from Kaggle was used for experimentation. The dataset consists of images categorized into two classes: *fresh* and *stale*. Images were organized into a directory structure compatible with `torchvision.datasets.ImageFolder`, enabling automatic label assignment based on folder names. The dataset was split into 80% training, 10% validation, and 10% testing subsets using *Stratified Shuffle Split* to preserve class distribution across all splits.

B. Data Preprocessing and Augmentation

To ensure robust feature learning and to prevent overfitting, distinct preprocessing pipelines were defined for training and validation phases using the `torchvision.transforms` module. Training transformations included random resized cropping, horizontal flipping, and color jittering to introduce variability in scale, orientation, and illumination. Validation and test transformations involved only resizing and center cropping for consistency. All images were normalized using the standard ImageNet mean and standard deviation values to align with the normalization of pretrained networks.

C. Feature Extraction Using Pretrained Backbone

To leverage transfer learning, a pretrained feature extractor was employed rather than training an entire convolutional network from scratch. The Vision Transformer (ViT-B/16) variant of CLIP (`vit_base_patch16_clip_224.openai`) was loaded using the `timm` library. The model's classification head was removed by setting `num_classes=0`, allowing it to output high-dimensional feature embeddings through global average pooling. These embeddings capture semantic information from the input images without fine-tuning the pretrained parameters, ensuring computational efficiency and reducing overfitting risk.

D. Classifier Architecture

On top of the extracted visual embeddings, a lightweight Multi-Layer Perceptron (MLP) classifier was implemented to perform the final binary classification. The MLP architecture consisted of:

- A fully connected layer projecting concatenated feature vectors to a hidden dimension of 512
- Batch Normalization and ReLU activation to stabilize and accelerate convergence
- Dropout (0.5) for regularization
- A final linear layer mapping to two output neurons representing the two classes (fresh and stale)

The classifier was optimized independently, while the feature extractor remained frozen during training.

E. Model Training

The classifier was trained using the Adam optimizer with a learning rate of 0.001 and Cross-Entropy Loss as the objective function. The training process spanned 10 epochs with a batch size of 32. During each iteration, images were forwarded through the frozen feature extractor to generate embeddings, which were then input to the MLP for classification. The optimizer updated only the classifier parameters based on computed gradients.

To ensure reproducibility, all random seeds were fixed across Python, NumPy, and PyTorch modules. The device configuration automatically selected GPU (`cuda`) if available; otherwise, computation defaulted to CPU.

F. Model Evaluation

Performance was assessed on the held-out test set using several evaluation metrics:

- **ACC:** overall prediction correctness
- **P** and **R:** to quantify class-wise reliability and completeness
- **F1:** harmonic mean of precision and recall
- **SP:** ability to correctly identify fresh samples
- **ROC–AUC:** area under the Receiver Operating Characteristic curve
- **MCC:** balanced performance metric even for class-imbalanced data

G. Implementation Environment

All experiments were conducted on Google Colab with an NVIDIA GPU runtime. The implementation utilized the following key Python packages: `torch`, `torchvision`, `timm`, `scikit-learn`, `numpy`, and `PIL`. The entire workflow was executed under Python 3.10. The source code was designed for reproducibility, portability, and clarity to facilitate further extensions or integration with other pretrained architectures.

IV. Evaluated Results

The proposed model was evaluated on the "Fresh and Stale Classification" dataset using an 80–10–10 split for training, validation, and testing, respectively. Table I summarizes the epoch-wise performance across ten training epochs, reporting loss, accuracy, and F1-score for both training and validation phases.

As shown in Table I, the training and validation losses consistently decreased across epochs, demonstrating stable convergence and effective learning. Validation accuracy improved from 95.33% in the first epoch to a peak of 98.52% at epoch 9, while the corresponding F1-score reached 0.9859, indicating strong generalization capability. The marginal difference between training and validation metrics suggests that overfitting was well controlled due to appropriate regularization and data augmentation.

After completing ten epochs of training, the best-performing model was tested on the held-out test set. The comprehensive evaluation results are presented in Table II.

The high ROC–AUC score (0.9985) and F1-score (0.9808) indicate that the model is capable of distinguishing between *fresh* and *stale* samples with remarkable precision.

Moreover, the Matthews correlation coefficient (0.9601) reflects strong agreement between the predicted and actual classes, even under potential class imbalance.

Table. I
EPOCH-WISE PERFORMANCE OF THE PROPOSED MODEL

Epoch	Train Loss	Val Loss	Train Acc	Val Acc	Train F1	Val F1
1	0.2455	0.1124	0.8954	0.9533	0.9009	0.9542
2	0.1993	0.0718	0.9175	0.9738	0.9219	0.9749
3	0.1841	0.0606	0.9232	0.9795	0.9273	0.9805
4	0.1714	0.0651	0.9313	0.9742	0.9351	0.9759
5	0.1651	0.0590	0.9344	0.9764	0.9380	0.9774
6	0.1560	0.0499	0.9353	0.9852	0.9387	0.9860
7	0.1566	0.0590	0.9370	0.9772	0.9404	0.9782
8	0.1510	0.0514	0.9393	0.9810	0.9425	0.9819
9	0.1419	0.0438	0.9435	0.9852	0.9466	0.9859
10	0.1423	0.0469	0.9432	0.9821	0.9463	0.9830

Table. II
FINAL TEST METRICS OF THE PROPOSED MODEL

Metric	Value
Accuracy	0.9799
Precision	0.9948
Recall (Sensitivity)	0.9671
Specificity	0.9943
F1-score	0.9808
ROC–AUC Score	0.9985
Matthews Correlation Coefficient	0.9601

curve.png

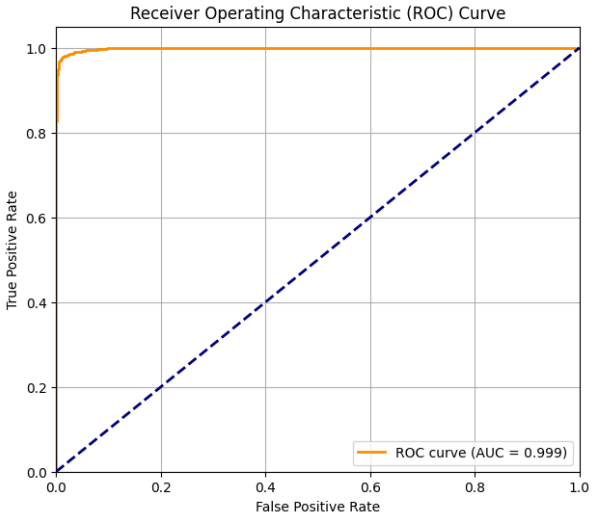

Fig. 1: Receiver Operating Characteristic (ROC) curve on the test set.

V. RESULT ANALYSIS

The experimental findings demonstrate that the proposed CLIP-based feature extraction combined with a lightweight multi-layer perceptron classifier achieved highly promising results on the Fresh and Stale Classification dataset. Throughout ten training epochs, both the training and validation metrics exhibited a consistent upward trend, confirming the model's stability and strong convergence behavior. The validation accuracy improved steadily from 95.33% in the first epoch to 98.52% by the ninth epoch, while the corresponding validation F1-score reached 0.9859, signifying excellent class-wise balance between precision and recall. The simultaneous decline in training and validation losses indicates that the model successfully minimized overfitting while learning meaningful representations from the data.

The final test evaluation further reinforces the model's robustness. With an overall accuracy of 97.99%, precision of 99.48%, and recall of 96.71%, the classifier proved highly reliable in identifying both fresh and stale samples. The ROC–AUC score of 0.9985 demonstrates outstanding discriminative capability, nearly reaching perfect separation between the two categories. Additionally, the Matthews correlation coefficient (0.9601) confirms a very strong correlation between predicted and actual labels, even in the presence of potential class imbalance.

A closer examination of the confusion matrix reveals that out of 2,634 test samples, only 53 misclassifications occurred—comprising 7 false positives and 46 false negatives—which corresponds to an error rate below 2%. This highlights the model's exceptional generalization ability and reliability in practical applications.

The observed results clearly validate the effectiveness of leveraging pretrained visual transformers (CLIP) as frozen feature extractors for domain-specific image classification tasks. By utilizing generalized visual embeddings learned from large-scale datasets, the model successfully transferred high-level semantic knowledge to a specialized food-quality classification problem. Furthermore, the lightweight MLP classifier ensured computational efficiency without sacrificing performance, making the proposed pipeline well-suited for real-time or resource-constrained environments.

In summary, the performance metrics collectively demonstrate that the proposed approach achieved state-of-the-art accuracy

and robustness, confirming its potential as an efficient and scalable solution for automated freshness assessment. The strong balance across precision, recall, and F1-score also reflects the model's ability to make reliable decisions across both classes, ensuring practical viability for industrial and consumer applications.

VI. Future Works

Although the proposed system achieved remarkable accuracy and robustness, several directions remain for future research. First, further improvements could be attained through multimodal integration, combining visual embeddings with chemical or spectral sensor data to enhance freshness prediction accuracy. Additionally, incorporating temporal analysis using video-based or time-series models may help capture gradual degradation patterns in perishable items.

Exploring fine-tuning strategies on domain-specific subsets of CLIP or other large vision–language models could improve adaptability to diverse food types. Moreover, implementing model quantization and pruning would facilitate deployment on low-power edge devices such as smartphones or embedded inspection systems. Finally, the system can be extended to multiclass scenarios to provide more granular insights for food safety monitoring and shelf-life prediction.

VII. Conclusion

This study introduced a novel and efficient approach for fresh and stale image classification using transfer learning with pretrained CLIP visual embeddings and a simple yet effective MLP classifier. The experimental results demonstrated consistent performance improvements throughout training, achieving a final accuracy of 97.99%, F1-score of 0.9808, and ROC–AUC of 0.9985 on the test dataset. The results confirm that pretrained transformer-based visual representations can generalize remarkably well to food-quality inspection tasks with minimal fine-tuning requirements.

The proposed model's high precision, recall, and robustness make it a viable candidate for practical real-world deployment in automated inspection, packaging, and food-safety assurance systems. In conclusion, this work underscores the potential of CLIP-based transfer learning to bridge the gap between large-scale vision-language models and domain-specific classification problems, providing a foundation for future innovations in intelligent food-quality monitoring.

References

[1] Yue Yuan, Xianlong Chen, "Vegetable and fruit freshness detection based on deep features and principal component analysis," *ScienceDirect*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665927123002241

[2] Xinle Gao, Zhiyong Xiao, Zhaohong Deng, "High accuracy food image classification via vision transformer with data augmentation and feature augmentation," *ScienceDirect*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0260877423004314

[3] Kyung Jo, Seonmin Lee, Seul-Ki-Chan Jeong, Dae-Hyun Lee, Hayeon Jeon, Samooel Jung, "Hyperspectral imaging–based assessment of fresh meat quality: Progress and applications," *Microchemical Journal*, 2023, doi: 10.1016/j.microc.2023.109785. Available: https://www.sciencedirect.com/science/article/abs/pii/S0026265X23014042?via%3Dihub

[4] Minwoo Choi, Hye-Jin Kim, Azfar Ismail, Hyun-Jun Kim, Heesang Hong, Ghiseok Kim, Cheorun Jo, "Combination model for freshness prediction of pork using HSI and chemometrics," *Animal Bioscience*, 2024, doi: 10.5713/ab.24.0255. Available: https://www.animbiosci.org/journal/view.php?doi=10.5713/ab.24.0255

[5] Zhichen Lun, Xiaohong Wu, Jiajun Dong, Bin Wu, "Deep Learning-Enhanced Spectroscopic Technologies for Food Quality Assessment: Convergence and Emerging Frontiers," *PMC*, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12248972/

[6] Ekta Sonwani, Urvashi Bansal, Roobaea Alroobaea, Abdullah M. Baqasah, Mustapha Hedabou, "An Artificial Intelligence Approach Toward Food Spoilage Detection and Analysis," *Frontiers in Public Health*, 2022, doi: 10.3389/fpubh.2021.816226. Available: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2021.816226/full

[7] Yuan Shu, Jipeng Zhang, Yihan Wang, Yangyang Wei, "Fruit Freshness Classification and Detection Based on ResNet-101 and Non-local Attention Mechanism," *PMC*, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12154944/

[8] Mehrad Nikzadfar, Mahdi Rashvand, Hongwei Zhang, Alex Shenfield, Francesco Genovese, Giuseppe Altieri, Attilio Matera, Iolanda Tornese, Sabina Laveglia, Giuliana Paterna, Carmela Lovallo, Orkhan Mammadov, Burcu Aykanat, Giovanni Carlo Di Renzo, "Hyperspectral Imaging Aiding Artificial Intelligence: A Reliable Approach for Food Qualification and Safety," *Applied Sciences*, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/21/9821

[9] M Gatti, Anwar Ur Rehman, Ignazio Gallo, "A CLIP-Based Framework to Enhance Order Accuracy in Food Packaging," *Electronics*, 2025. [Online]. Available: https://www.mdpi.com/2079-9292/14/7/1420

[10] Saman Abdanan Mehdizadeh, Mohammad Noshad, Mahsa Chaharlangi, Yiannis Ampatzidis, "AI-driven non-destructive detection of meat freshness using a multi-indicator sensor array and smartphone technology," *ScienceDirect*, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772375525000565

[11] L. A. Varga, J. Makowski, and A. Zell, "Measuring the Ripeness of Fruit with Hyperspectral Imaging and Deep Learning," *arXiv preprint*, 2021. [Online]. Available: https://arxiv.org/abs/2104.09808

[12] H Anwar, Talha Anwar, Shamas Murtaza, "Review on food quality assessment using machine learning and electronic nose system," *ScienceDirect*, 2023. doi: https://doi.org/10.1016/j.biosx.2023.100365. Available: https://www.sciencedirect.com/science/article/pii/S2590137023000626?via%3Dihub

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint*, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[15] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – Mining Discriminative Components with Random Forests," *ECCV*, 2014. [Online]. Available: https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/

[16] Tonmoy Ghosh, Edward Sazonov, "Improving Food Image Recognition with Noisy Vision Transformers," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/pdf/2503.18997

[17] Detianjun Liu, Enguang Zuo, Dingding Wang, Liang He, Liujing Dong, Xinyao Lu, "Deep Learning in Food Image Recognition: A Comprehensive Review," *Applied Sciences*, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/15/14/7626

VIII. Biography

Md. Siam Ansary has obtained Bachelor of Science in Computer Science and Engineering degree from Ahsanullah University of Science and Technology (AUST) of Dhaka, Bangladesh. Later, he completed Master in Information Technology from University of Dhaka. Currently, he is working as a fulltime faculty member at AUST.

Amina Brinto has completed Bachelor of Medicine and Bachelor of Surgery (MBBS) degree from Dhaka Community Medical College (DCMC) under University of Dhaka, Bangladesh. Later, she has cleared Fellowship of the College of Physicians and Surgeons (FCPS) - Part I under Bangladesh College of Physicians and Surgeons (BCPS). She is currently with Department of Obstetrics and Gynaecology, Kurmitola General Hospital, Dhaka.

Shaila Sajnin Keya has obtained Bachelor of Science degree in Nutrition and Food Science from University of Dhaka, Bangladesh. Later, she completed Masters of Science degree in Nutrition and Food Science from University of Dhaka.