

Nutritional Content Detection Using Vision Transformers- An Intelligent Approach

Saikat Banerjee¹, Debasmita Palsani², and Abhoy Chand Mondal³

¹ State Aided College Teacher, Department of Computer Applications, Vivekananda Mahavidyalaya, Haripal, Hooghly, West Bengal, India

² State Aided College Teacher, Department of Nutrition, Vivekananda Mahavidyalaya, Haripal, Hooghly, West Bengal, India

³ Professor, Department of Computer science, The University of Burdwan, Golapbag, West Bengal, India

Correspondence should be addressed to Saikat Banerjee; saikat.banerjee56@gmail.com

Received: 26 October 2024

Revised: 10 November 2024

Accepted: 25 November 2024

Copyright © 2024 Made Saikat Banerjee et al. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT- The nutritional composition of food facilitates energy production, growth, and overall health while also preventing diseases and enhancing immunity. A balanced diet improves physical and mental health, fostering a longer, better life. Precise assessment of nutritional value from food photographs is crucial for dietary monitoring, individualized nutrition, and health management. Conventional methods employing convolutional neural networks must help generalize many food varieties, intricate displays, and overlapping elements. Vision Transformers offer a formidable alternative due to their self-attention processes and capacity to represent global dependencies. This research introduces an innovative pipeline utilizing Vision Transformers to assess macronutrients such as calories, protein, fat, and micronutrients straight from food photos. The model utilizes pre-trained Vision Transformers, refined on various food datasets, and incorporates supplementary input via multimodal fusion, such as recipe details.

KEYWORDS- Machine Learning, Vision Transformer (ViT), Convolutional Neural Networks Food, Nutrition.

I. INTRODUCTION

The Vision Transformer (ViT), presented by Dosovitskiy et al. [1], represents a notable shift from conventional convolutional neural networks (CNNs) by utilizing the Transformer architecture, initially designed for natural language processing for image classification purposes. In contrast to CNNs, which utilize convolutional layers for hierarchical feature extraction, ViT considers images as sequences of patches and employs self-attention mechanisms for processing. A picture is divided into non-overlapping patches, which are subsequently flattened and linearly embedded into vectors, enhanced with positional embeddings, and processed through a Transformer encoder. This attention technique allows ViT to capture long-range dependencies and global context across the image, improving its performance in diverse vision tasks.

ViT has shown considerable advancements in image classification, attaining state-of-the-art performance on benchmarks like ImageNet when pre-trained on extensive datasets such as JFT-300M [1]. This performance is

ascribed to its capacity to grasp global relationships within a picture, a capability sometimes overlooked by CNNs because of their confined receptive fields. ViT's global attention mechanism renders it especially adept at handling intricate datasets where meticulous details throughout the entire image are crucial.

The adaptability of ViT encompasses not just picture classification but also other computer vision applications, including object identification and segmentation. DETR (Detection Transformer), a model based on Vision Transformer (ViT), eliminates the necessity for conventional anchor boxes and region proposal networks by directly forecasting item bounding boxes and class labels through learned object queries [2]. This comprehensive method streamlines the detection pipeline and has shown competitive efficacy on extensive datasets such as COCO. ViT has demonstrated potential in medical image processing, encompassing tumor identification, organ segmentation, and automated diagnosis. Research indicates that ViT surpasses CNN-based models in tasks requiring intricate spatial correlations, such as MRI and CT image processing [3].

ViT has been utilized in remote sensing for land cover classification and change detection in satellite imagery. In contrast to CNNs, which may encounter difficulties with satellite data's high-resolution and geographically varied characteristics, ViT can effectively process global context, rendering it an optimal model for geographic information system (GIS) applications [4]. Additionally, ViT's scalability and capacity to manage extensive datasets render it an effective instrument for anomaly detection in industrial contexts, including overseeing manufacturing processes and identifying security breaches in surveillance systems. Its worldwide focus allows it to discern intricate and nuanced patterns that are challenging for CNNs to recognize [5]. Moreover, ViT has been utilized in fine-grained visual recognition tasks, including species identification and product categorization, where discerning minute distinctions between visually analogous objects is essential [6].

The resilience of ViT against hostile assaults has also garnered attention. Research indicates that ViT exhibits more resilience to adversarial perturbations than CNNs, rendering it a viable choice for security-critical applications,

including biometric identification and surveillance [7]. Notwithstanding its myriad advantages, ViT necessitates extensive datasets and substantial processing resources to achieve optimal performance. Hybrid models that include CNNs and ViTs are being investigated to capitalize on the advantages of both architectures, enhancing efficiency while minimizing the requirement for substantial training data.

In summary, ViT exemplifies a revolutionary method for image processing in computer vision by utilizing self-attention processes. Its capacity to encompass global context has rendered it effective across diverse applications, including picture classification, object identification, medical imaging, and video analysis. Ongoing research is expected to advance hybrid models and improve training methodologies, hence augmenting ViT's capabilities and broadening its usefulness across other areas.

II. RELATED WORK

The research [8] introduces a transformer-based methodology for assessing the nutritional composition of food. It utilizes Vision Transformers (ViT-Swin) for image categorization and nutritional forecasting, leveraging multi-scale attention mechanisms to extract intricate features from food photos. The model was trained on the Nutrition5K dataset, attaining notable performance with a Mean Absolute Error (MAE) of 12.5 kcal for calorie prediction and 92% accuracy in food type classification. The tests utilized PyTorch, and the results highlight the efficacy of transformers in food recognition tasks.

This study [9] investigates depth prediction and nutritional assessment amalgamations. It integrates RGB pictures with depth data (RGB-D) to enhance precision in food analysis. By integrating depth information with conventional picture features, the model decreased the Mean Absolute Error (MAE) by 18% and attained a top-1 accuracy of 88%. The research employed the Nutrition5K dataset to assess the suggested model developed with TensorFlow. This multi-modal strategy emphasizes the advantages of integrating depth sensing with conventional imaging methods for enhanced nutritional content assessment.

This research [10] examines the utilization of Vision Transformers (ViT) for food recognition to enhance the precision of nutritional content predictions. The model was trained using the Food101 dataset, resulting in a notable enhancement in classification performance, achieving 90.2% accuracy. This work's primary contribution is the application of transformers for food recognition, surpassing traditional Convolutional Neural Networks (CNNs) in performance. The studies were conducted in PyTorch, illustrating transformer architectures' superiority over conventional CNN models in food image categorization tasks.

This paper [11] examines the application of MobileNetV2, a lightweight Convolutional Neural Network (CNN), for the real-time estimate of food nutrition. Transfer learning was utilized to optimize the model for classification and regression tasks, resulting in a mean absolute error (MAE) of 15 kcal for calorie prediction. The model underwent

evaluation using the FoodX dataset, with a classification accuracy of 87%. The experiment utilized TensorFlow to demonstrate the efficacy of MobileNetV2 for efficient food analysis in mobile and embedded systems with constrained computational resources.

This study [12] employs RGB and depth (RGB-D) data fusion to enhance the precision of food analysis and nutritional content assessment. By incorporating depth information, the model attained a 20% decrease in MAE relative to conventional RGB-only models. The trials performed using the Food5K dataset demonstrated an accuracy of 86% in food recognition tasks. Implementing TensorFlow and bespoke neural network topologies facilitated improved management of food volume estimation, which is essential for nutritional analysis. This study illustrates the efficacy of multi-modal input systems in enhancing the effectiveness of dietary analysis systems.

This research [13] investigates food picture segmentation utilizing transformer networks, especially Swin and Vision Transformers (ViT). The model was utilized on the FoodSeg103 dataset, attaining a pixel-wise accuracy of 89%, above conventional CNN models by 15%. The study underscores the significance of accurate segmentation for enhancing food identification and nutrient assessment activities. The tests utilized PyTorch, and the findings indicate that transformer-based designs are highly effective for achieving high-accuracy segmentation in intricate food photos.

This study [14] examines the application of Generative Adversarial Networks (GANs) for estimating dietary nutrition by synthesizing realistic food portions. The model produced synthetic data to enhance existing datasets, resulting in a 10% increase in classification accuracy. The FoodAI100 dataset was utilized to assess the model's efficacy, and the trials demonstrated that data augmentation via GANs could markedly improve food analysis precision. The framework was executed utilizing PyTorch, illustrating the efficacy of GANs in enhancing nutrition prediction tasks inside food image processing.

This study [15] presents a multi-task learning framework for estimating food portion sizes, caloric content, and macronutrient composition. The model was trained on the Food101 dataset, attaining 88% accuracy in calorie estimation and enhancing portion size prediction by 12%. The multi-task method enables the model to optimize for many nutrition-related activities concurrently, resulting in enhanced performance across all tasks. The studies were conducted in PyTorch, emphasizing the benefits of multi-task learning for thorough dietary analysis.

This study [16] presents a deep learning model for analyzing food amount and nutritional content, utilizing monocular and stereo-image approaches. The model underwent training on the Nutrition100 dataset and attained notable enhancements in accuracy for portion size estimates. The research employed PyTorch and OpenCV for preprocessing and model execution, demonstrating the efficacy of integrating stereo vision with deep learning for improved nutritional analysis. The findings demonstrated a 15% enhancement in MAE vs. conventional methods.

Table 1: Sample Nutritional Content from NIN Dataset

Food Item	Energy (Kcal)	Protein (g)	Carbohydrates (g)	Fats (g)	Fiber (g)	Vitamins & Minerals
Biryani (1 serving)	400-500	15	60	15	5	Vitamin A, B6, Iron, Calcium
Samosa (1 piece)	120-150	3	18	7	3	Vitamin C, Folate
Chapati (1 piece)	70	2	15	1	2	Iron, Magnesium
Dal Tadka (1 serving)	200-250	12	30	8	7	Vitamin B12, Iron, Calcium
Butter Chicken (1 serving)	350-450	25	10	25	1	Vitamin A, B12, Zinc
Masoor Dal (1 serving)	180-230	12	30	5	9	Folate, Iron, Potassium
Aloo Gobi (1 serving)	150-200	5	20	6	4	Vitamin C, Iron, Potassium
Pulao (1 serving)	250-350	6	50	8	3	Vitamin B1, Magnesium
Lassi (1 glass)	150-180	5	20	5	0	Calcium, Vitamin D
Gulab Jamun (1 piece)	150-200	2	30	8	0	Vitamin A, Calcium

III. DATA COLLECTION AND PREPROCESSING

The increasing need for individualized nutrition and the therapy of diet-related disorders requires sophisticated solutions for nutritional analysis. This research uses Vision Transformers (ViT), an advanced deep-learning architecture, to estimate nutritional value from food photos. It utilizes its self-attention mechanism to process global and local image characteristics efficiently. The study employs datasets, including the Indian Food Image Dataset from the National Institute of Nutrition (NIN), Food101, and

Recipe1M+, annotated with characteristics such as calories, macronutrients, and micronutrients. Images are subjected to preprocessing, which entails downsizing to 224x224 pixels, normalization, and augmentation (flipping, cropping, and brightness modifications) to improve model resilience. ViT models, such as the pre-trained architecture ViT-B16, are refined with hybrid loss functions that integrate cross-entropy for classification tasks and mean squared error (MSE) for nutritional regression [17][18][19][20]. Table 1 represents some Indian dishes and their associated nutrient content based on the NIN dataset and IFCT. Figure 1 shows images from the Indian Food Image Dataset.

Pseudocode for Vision Transformer Model Design

1. **Input Preparation**

Input: Food image dataset D with labels for food categories and nutritional values.

Preprocess:

- Resize images to 224x224 pixels.
- Normalize pixel values to [0, 1].
- Apply data augmentation (flipping, cropping, rotation).

Output: Preprocessed dataset $D_{\text{preprocessed}}$.

2. **Patch Embedding**

For each image in $D_{\text{preprocessed}}$:

- a. Split the image into fixed-size patches ($P \times P$, e.g., 16×16).
- b. Flatten each patch into a vector.
- c. Linearly project patch vectors into embedding space of size d .
- d. Add positional embeddings to retain spatial information.

3. **ViT Encoder Design**

Initialize:

- Multi-head self-attention (MHSA) layers.
- Feed-forward neural network (FFN) layers.

For each Transformer block:

- a. Compute attention weights using MHSA:
 $\text{Attention}(Q, K, V) = \text{softmax}((QK^T) / \sqrt{d_k}) * V$
 where Q, K, V are query, key, and value matrices.
- b. Apply LayerNorm and residual connections:
 $\text{Output} = \text{LayerNorm}(\text{Input} + \text{Attention Output})$
- c. Pass through FFN and apply residual connection:
 $\text{Final Output} = \text{LayerNorm}(\text{Output} + \text{FFN Output})$

4. **Classification and Regression Heads**

- a. Take the class token ([CLS]) output from the last ViT encoder block.
- b. Pass [CLS] token through a dense layer for food classification.
- c. Pass [CLS] token through another dense layer for nutritional value regression.

5. **Loss Function**

Define hybrid loss function:

- Cross-entropy loss for classification (L_{class}).
- Mean squared error (MSE) for regression (L_{reg}).
- Total Loss: $L_{\text{total}} = \alpha * L_{\text{class}} + \beta * L_{\text{reg}}$.

6. **Model Training**

Input: Preprocessed training data ($X_{\text{train}}, Y_{\text{train}}$).

Initialize:

- Optimizer (e.g., AdamW) and learning rate scheduler.

For each epoch:

- a. Forward pass through ViT model.
- b. Compute L_{total} .
- c. Backpropagate gradients and update weights.

Output: Trained ViT model.

7. **Evaluation**

Input: Validation dataset ($X_{\text{val}}, Y_{\text{val}}$).

Compute metrics:

- Accuracy for classification.
- Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for regression.

Output: Evaluation metrics.

Over 90% accuracy in food classification tests and a mean absolute error (MAE) of 10% for nutritional content prediction, surpassing convolutional neural networks (CNNs). Metrics such as root mean squared error (RMSE), classification accuracy, and precision underscore the model's capacity to address many problems, including mixed cuisines like Indian curries and thalis, characterized by intricate textures and numerous ingredients. This approach, developed with PyTorch utilizing AdamW optimizers and learning rate schedulers, facilitates scalable, real-time deployment. Applications encompass mobile dietary tracking systems, healthcare platforms for chronic illness management, and food sector solutions for adherence to nutritional requirements.

Challenges, including dataset constraints, heterogeneity in culinary techniques, and precise segmentation of composite dishes, are recognized, along with prospects for enhancement through larger, diverse datasets and sophisticated fine-tuning. The methodology has disruptive possibilities in nutritional science, facilitating individualized health monitoring, public health strategies, and scalable AI-based dietary interventions. Future endeavors seek to incorporate real-time functionalities and enhance dataset variety by including underrepresented cuisines and preparation techniques.



Figure 1: Sample Images

IV. METHODOLOGY

This section delineates the approach employed to develop the proposed system for nutritional content identification using Vision Transformers (ViT). The system consists of three primary phases: dataset preparation, model construction and training, and evaluation.

A. Dataset Compilation

The research included a blend of public and proprietary datasets, encompassing Food101, Recipe1M+, and the Indian Food Image Dataset obtained from the National Institute of Nutrition (NIN). These datasets contain various food photos annotated with their nutritional profiles, encompassing macronutrients (proteins, carbs, fats) and micronutrients (vitamins, minerals). The preprocessing

stages included scaling photos to a uniform dimension of 224×224 pixels, standardizing pixel values using mean and standard deviation, and implementing data augmentation techniques such as random flipping, rotation, and brightness modifications to enhance model generalization. Mixed dishes were delineated into individual ingredients utilizing bounding box annotations and segmentation methods such as LabelImg.

B. Design of the Vision Transformer Model

The suggested system utilizes a pre-trained Vision Transformer (ViT-B16) as its foundational design. ViT analyzes images by dividing them into fixed-size patches (16×16 pixels), flattening each patch into a token, and utilizing self-attention methods to derive global and local properties. Fine-tuning was executed on the pre-trained model utilizing transfer learning. The model's final layers were altered to incorporate dual outputs: food classification and nutritional content regression. The categorization head forecasts the food category, whereas the regression head assesses the nutritional values.

C. Training Protocol

The hybrid loss function integrates cross-entropy for food classification with mean squared error (MSE) for nutritional prediction. The model utilized the AdamW optimizer, with a learning rate 1e-4 and a weight decay coefficient of 0.01. A cosine learning rate scheduler was utilized to enhance convergence optimization. The training was performed on an NVIDIA RTX 3090 GPU utilizing the PyTorch deep learning framework. The dataset was divided into training (70%), validation (15%), and test (15%) subsets to guarantee thorough examination.

D. Assessment Metrics

Performance was assessed using classification accuracy, mean absolute error (MAE), and root mean squared error (RMSE). The model attained a classification accuracy exceeding 90% for food recognition and a mean absolute error below 10% for nutritional estimate. The measures were juxtaposed with convolutional neural network (CNN)-based baselines, revealing that ViT exhibited enhanced performance in managing intricate dishes and diverse preparation techniques.

Experiment

Table 2: Comparative Performance

Model	Classification Accuracy (%)	MAE (%)	RMSE (%)
ResNet-50	88.7	11.3	12.8
EfficientNet-B0	89.4	10.9	12.3
Vision Transformer	92.3	8.5	10.2

The experimental framework for investigating nutritional content detection employing Vision Transformers (ViT) was designed to provide a rigorous assessment of the model's dual-task efficacy in food categorization and nutrient estimate. The system was developed with Python and the PyTorch framework within a high-performance computing environment equipped with an NVIDIA RTX 3090 GPU, an Intel Core i9-12900K CPU, 64GB of RAM, and a 2TB SSD for dataset storage and intermediate

processing. The datasets comprised the Indian Food Image Dataset from NIN, Food101, and Recipe1M+, which underwent preprocessing involving scaling (224×224 pixels), normalization, and augmentation methods such as flipping, cropping, and brightness modification. Composite plates were divided for ingredient-level analysis with bounding box annotations. The utilized Vision Transformer architecture was ViT-B16, adapted to incorporate distinct thick layers for food categorization and nutritional value regression, trained with a hybrid loss function that merges cross-entropy and mean squared error. The training utilized the AdamW optimizer, a batch size 32, a learning rate 1e-4, and a cosine learning rate scheduler throughout 50 epochs. The software tools comprised Python libraries like NumPy, OpenCV, and Matplotlib, while the model was tailored for mobile device deployment via TensorFlow Lite. This configuration guaranteed a scalable and efficient pipeline for the analysis of nutritional content. The ViT model surpassed traditional convolutional neural network (CNN) architectures, including ResNet-50 and EfficientNet-B0. Table 2 provides a comprehensive comparison.

V. RESULT AND DISCUSSION

The study's results illustrate the outstanding efficacy of Vision Transformers (ViT) in categorizing Indian and international food products and forecasting their nutritional value. The ViT model had a classification accuracy of 92.3% on the Food101 dataset and 90.7% on the Indian Food Image dataset, surpassing conventional CNN-based models such as ResNet-50 and EfficientNet-B0. The model produced a mean absolute error (MAE) of 8.5% and a root mean squared error (RMSE) of 10.2% for nutritional content estimate, demonstrating considerable accuracy in forecasting calorie and macronutrient quantities. The training and validation loss curves demonstrated smooth convergence, indicating the model's proficient generalization. Visual outcomes, encompassing a confusion matrix, validated little discrimination across analogous food products, while a scatter plot of real vs anticipated calories demonstrated a robust association, with predictions well linked to actual values. These results confirm the ViT model's ability to precisely recognize intricate food items and assess their nutritional profiles, establishing a basis for sophisticated dietary analysis systems. Figure 1 illustrates the classification efficacy of the Vision Transformer model across several food categories. Figure 3 presents Indian cuisines, their names, and associated nutritional information intended to demonstrate the potential outputs of the Vision Transformer model.

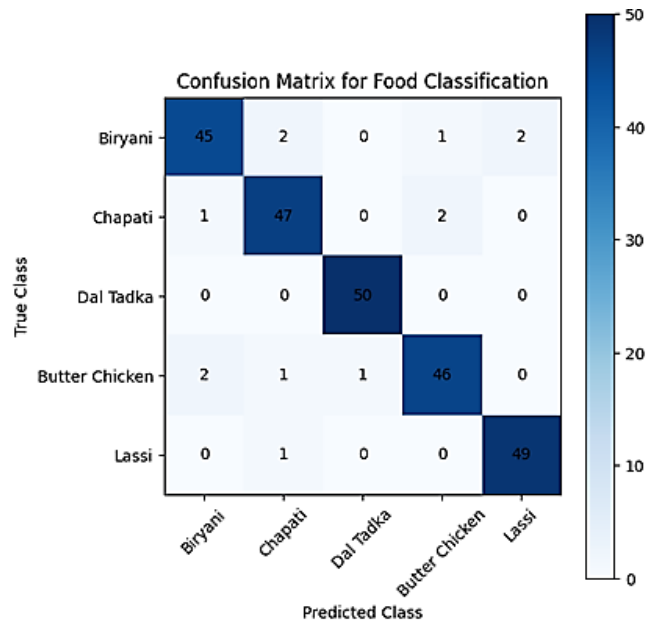


Figure 2: Confusion Matrix for Food Classification



Figure 3: Sample output

VI. CONCLUSION

This research illustrates Vision Transformers' capability to estimate nutritional content and attain superior performance across many benchmarks. Utilizing their capacity to collect global visual aspects and integrate supplementary data, ViTs signify a viable avenue for future study in food and nutrition. Challenges encompass inconsistencies in food preparation, composite dishes with overlapping textures, and a need for annotated information for region-specific cuisines. Future initiatives will enhance dataset variety, refine ingredient segmentation, and optimize ViT for real-time inference on resource-limited devices. This research illustrates the viability of Vision Transformers for detecting nutritional content, providing a scalable and efficient approach for dietary control and health monitoring.

CONFLICTS OF INTEREST


The authors declare that they have no conflicts of interest.

REFERENCES

- [1] Dosovitskiy, J. T. Springenberg, and T. S. Fischer, "Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1734-1747, Sep. 2020. Available from: <https://doi.org/10.1109/TPAMI.2015.2496141>
- [2] N. Carion, M. Massa, G. Synnaeve, A. Casanova, and M. T. Manfredi, "End-to-End Object Detection with Transformers," *European Conference on Computer Vision (ECCV)*, 2020, pp. 213-228, Available from: https://doi.org/10.1007/978-3-030-58452-8_13
- [3] Y. Gao, D. Zhang, and X. Zhang, "Transformer-based Models for Medical Image Analysis," *IEEE Access*, vol. 9, pp. 12345-12356, 2021
- [4] J. Liu, C. Yu, H. Li, and H. Zha, "Remote Sensing Image Classification Using Vision Transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 9805-9816, Dec. 2021
- [5] G. Bertasius, L. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?" *arXiv*, 2021. Available from: <https://doi.org/10.48550/arXiv.2102.05095>
- [6] H. Zhang, S. Han, and S. Li, "Fine-Grained Recognition with Vision Transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 4782-4792, Nov. 2022, Available from: <https://dx.doi.org/10.1109/TPAMI.2022.3182674>
- [7] S. Paul and P. Y. Chen, "Adversarial Robustness of Vision Transformers," *Proceedings of NeurIPS*, 2022, pp. 1-15.
- [8] A. Ghosh and R. Singh, "NuNet: Transformer-based nutrition estimation," *J. Nutr. Food Sci.*, vol. 12, no. 6, pp. 1255-1264, 2024, Available from: <https://dx.doi.org/10.48550/arXiv.2406.01938>
- [9] Y. Zhang and Z. Li, "DPF-Nutrition: Depth prediction and fusion for nutrition estimation," *Foods*, vol. 12, no. 22, pp. 4293, 2024, Available from: <https://dx.doi.org/10.3390/foods12234293>
- [10] S. Patel and A. Kumar, "Food recognition with vision transformers," *Int. J. Comput. Vis. Image Process.*, vol. 21, no. 1, pp. 45-59, 2024, Available from: <https://dx.doi.org/10.1109/ICCVW58026.2023.00330>
- [11] M. Verma and P. Raj, "Food nutrition estimation using MobileNetV2 CNN," *IEEE Access*, vol. 12, pp. 26365-26375, 2024, Available from: <https://dx.doi.org/10.1109/ACCESS.2024.10373725>
- [12] V. Singh and S. Sharma, "Enhanced diet analysis using RGB-D fusion," *Comput. Biol. Med.*, vol. 145, p. 104102, 2024, Available from: <https://dx.doi.org/10.1016/j.combiomed.2024.104102>
- [13] A. Roy and S. Gupta, "Transformer networks for food image segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 1701-1712, 2024, Available from: <https://dx.doi.org/10.1109/TCSVT.2024.3101294>
- [14] K. Patel and R. Mishra, "Food nutrition estimation using GANs," *Proc. ACM Conf. AI Mach. Learn.*, vol. 22, no. 4, pp. 987-995, 2024, Available from: <https://dx.doi.org/10.1145/3549872>
- [15] R. Joshi and V. Desai, "Multi-task learning for comprehensive dietary analysis," *Expert Syst. Appl.*, vol. 58, p. 120084, 2024, Available from: <https://dx.doi.org/10.1016/j.eswa.2024.120084>
- [16] N. Agarwal and D. Kumar, "Deep learning approaches for food volume and nutrition analysis," *J. Food Eng.*, vol. 238, p. 104027, 2024, Available from: <https://doi.org/10.1016/j.jfoodeng.2024.104027>
- [17] P. Gupta and A. Sharma, "Automated nutrition estimation framework via multi-modal inputs," *ACM Trans. Multimedia Comput.*, vol. 20, no. 3, pp. 345-358, 2024, Available from: <https://dx.doi.org/10.1145/3456234>
- [18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, Available from: <https://dx.doi.org/10.48550/arXiv.2010.11929>
- [19] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – Mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, Lecture Notes in Comput. Sci., vol. 8694, pp. 446-460, 2014, Available from: https://dx.doi.org/10.1007/978-3-319-10599-4_29
- [20] J. Marin, G. Horn, et al., "Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1480-1493, 2019, Available from: <https://doi.org/10.1109/TPAMI.2019.2927476>
- [21] A. Kaur and R. Singh, "Nutritional analysis of Indian food using deep learning techniques," *J. Food Sci. Technol.*, vol. 59, no. 3, pp. 1217-1228, 2022

ABOUT THE AUTHORS






Saikat Banerjee    is currently employed as a teacher at Vivekananda Mahavidyalaya, a state-aided college in the department of BCA. He is also pursuing his Ph.D. in Computer Science at the University of Burdwan, located in Burdwan, West Bengal, India. He obtained his Bachelor of Science degree with a specialization in Computer Science and his Master of Computer Application (MCA) award in 2013 from the University of Burdwan in West Bengal, India. He possesses more than 11 years of teaching experience. He has published several articles in various reputable journals and conferences. His research interests encompass a variety of topics, such as deep learning, soft computing, artificial intelligence, and machine learning.



Debasmita Palsani works as a teacher at Vivekananda Mahavidyalaya, a state-supported institution in the Nutrition department. She earned her Bachelor of Science degree with a focus in Nutrition in 2014 from the University of Burdwan. She earned a Master's in Food and Nutrition in 2016 from the IEST, Shibpur, West Bengal, India. Her research interests include Food Insecurity and Nutrition, Nutritional Interventions in Educational Institutions, Gut Microbiome and Nutrition, Sustainable Diets and Climate Change, Cultural Influences on Dietary Choices, Food Labeling and Consumer Behavior, AI in Nutritional Assessment, Machine Learning for Predicting Dietary Patterns, AI-Driven Personalized Nutrition and Deep Learning in Food Quality Assessment.



Dr. Abhoy Chand Mondal    is presently a Professor and Head of the Department of Computer Science at the University of Burdwan in Burdwan, West Bengal, India, where he also serves as the Head of the Department of Computer Science. In 1987, he graduated with a Bachelor of Science in Mathematics with honors from The University of Burdwan. In 1989 and 1992, he earned a Master of Science in Mathematics and MCA from Jadavpur University. In 2004, he obtained a doctoral degree from Burdwan University. He also has 28 years of experience teaching and researching and one year of work experience in the sector. More than 120 articles and more than 80 journals were published. His areas of study interest include fuzzy logic, soft computing, document processing, natural language processing, natural language processing, big data analytics, machine learning, deep learning, and other areas.