International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/https://doi.org/10.22362/ijcert/2024/v11/i1/v11i1s03
Volume 11, Special Issue 1, December 2024

Research Paper

Deep Learning-Based Food Image Classification Using Enhanced CNN Architecture and Data Augmentation Techniques

[1] D. Koteshwar Rao, [2] K. Sahiti, [3] J. Vamshi, [4] K.Uday Kiran, [5*] T. Srikar

[1] Assistant Professor, Department of CSE, St.Mary's Engineering College, Hyderabad, India

[2,3,4,5] B.Tech Student, Department of CSE (CyberSecurity), St.Mary's Engineering College, Hyderabad, India

**Corresponding Author(s): srikarbittu57@gmail.com*

Received: 25/09/2024 Revised: 16/10/2024 Accepted: 09/12/2024 Published: 31/12/2024

Abstract: Automated food image classification has emerged as a critical research area due to its applications in dietary monitoring, calorie estimation, and personalized nutrition. However, the variability in food appearances, including high intra-class variability and inter-class similarity, poses significant challenges. This study proposes an optimized deep learning-based solution using a custom convolutional neural network (CNN) architecture to address these challenges. The model incorporates advanced techniques such as batch normalization, dropout, and data augmentation to improve generalization and mitigate overfitting. The proposed model was evaluated on the widely used Food-101 dataset, consisting of over 100,000 images across 101 food categories. The images were preprocessed through resizing, normalization, and data augmentation to enhance diversity and robustness. The model was trained using the Adam optimizer with a learning rate of 0.001, and its performance was assessed using accuracy, precision, recall, and F1-score metrics. Results indicate that the proposed model achieved an accuracy of 93.2%, outperforming benchmark architectures such as VGG-16 (87.6%) and ResNet-50 (91.0%). It also demonstrated a balanced performance with a precision of 92.4%, recall of 93.1%, and an F1-score of 92.7%. These results underscore the effectiveness of the proposed architecture in capturing complex visual features inherent in food images. While the model performs exceptionally well, challenges remain in classifying visually similar food categories. Future work aims to address these limitations by integrating attention mechanisms and multi-modal data to enhance classification accuracy and real-world applicability. This research contributes to advancing the field of food image classification with robust and scalable solutions.

Keywords: Food Image Classification, Deep Learning, Convolutional Neural Network (CNN), Food-101 Dataset, Data Augmentation, Dietary Monitoring

---

1. Introduction

The rapid evolution of artificial intelligence (AI) and its applications in computer vision have brought unprecedented opportunities for automating tasks that were traditionally labour-intensive. One such task, food classification, has garnered significant attention due to its applications in various fields such as health monitoring, personalized nutrition, and culinary management. [1] Automated food image classification can serve as the backbone of modern dietary assessment tools, enabling individuals and healthcare providers to monitor food intake efficiently. Despite its potential, the task presents unique challenges that require sophisticated solutions.

Food images are inherently complex due to the high intra-class variability and inter-class similarity of food items. [2] For example, the same dish may appear differently depending on preparation style, lighting conditions, or plating, while visually similar foods such as spaghetti and noodles might belong to entirely different categories. Moreover, the presence of side dishes, condiments, and overlapping items further complicates the classification process. These factors make food classification fundamentally different from conventional image recognition tasks, such as classifying objects or animals. Traditional machine learning approaches to food image classification relied on hand-crafted features such as colour histograms, texture descriptors, and shape-based

methods. [3] While these methods provided valuable insights, they struggled to generalize across diverse datasets, especially when confronted with the variability seen in real-world food images. Moreover, their reliance on manual feature engineering limited scalability and adaptability to new data.

Deep learning, particularly convolutional neural networks (CNNs), has revolutionized the field of image recognition by automatically learning hierarchical features from raw image data. [4] Unlike traditional methods, CNNs eliminate the need for manual feature extraction, allowing models to adapt dynamically to complex datasets. They have shown exceptional performance across various domains, including medical imaging, autonomous driving, and facial recognition. [5] This success has spurred interest in applying deep learning techniques to food image classification, where their ability to handle large-scale data and capture intricate visual patterns offers significant advantages. The primary objective of this research is to explore the capabilities of CNNs in addressing the challenges of automated food image classification. Specifically, this study proposes an optimized CNN architecture tailored to the unique requirements of food classification tasks. By leveraging state-of-the-art techniques such as data augmentation, dropout, and batch normalization, the proposed model aims to enhance classification accuracy while mitigating common issues such as overfitting. [6]

This paper is organized as follows: Section 2 reviews related work in food image classification and the application of deep learning techniques in this domain. Section 3 outlines the dataset, preprocessing steps, and model architecture used in this study. Section 4 presents the experimental results, comparing the proposed approach with existing benchmarks. Section 5 discusses the implications of the findings and potential areas for improvement. Finally, Section 6 concludes with a summary of contributions and directions for future research. By addressing the unique challenges of food image classification, this study aims to contribute to the growing body of knowledge in computer vision and AI, paving the way for practical applications in dietary assessment and beyond.

Key Contributions: This study presents a comprehensive approach to automated food image classification using deep learning, making several notable contributions to the field:

- **Optimized CNN Architecture for Food Classification:** The study introduces a custom convolutional neural network (CNN) architecture specifically designed for the complexities of food image classification. By leveraging advanced techniques like dropout, batch normalization, and data augmentation, the model effectively addresses challenges such as high intra-class variability and inter-class similarity.

- **Achieving State-of-the-Art Performance:** Our proposed model achieves a classification accuracy of 93.2% on the Food-101 dataset, outperforming well-established architectures like VGG-16 and ResNet-50. This demonstrates the robustness and efficacy of the design in handling diverse and complex food datasets.

- **Foundation for Practical Applications:** By tackling key challenges in food image classification, this work provides a scalable and efficient solution with practical implications for real-world applications such as dietary monitoring, calorie estimation, and mobile-based food recognition systems.

2. Literature Review

The task of food image classification has evolved significantly over the years, driven by advancements in computer vision and machine learning. This section provides an overview of traditional approaches, highlights the transition to deep learning methods, and discusses recent advancements in the domain.

2.1 Traditional Approaches to Food Image Classification

Traditional machine learning methods for food image classification relied heavily on manually engineered features to capture the visual properties of food items. These approaches often employed:

Color-Based Features: Food items, often rich in visual texture, were distinguished using color histograms and other color-based metrics. For example, dishes like salads and soups were classified based on the predominance of green or liquid-like textures. [7]

Texture-Based Features: Texture descriptors such as the Local Binary Pattern (LBP) or Histogram of Oriented Gradients (HOG) were employed to capture the surface variations in food images. [8]

Shape-Based Features: Geometric and structural characteristics were used to differentiate foods with distinct outlines, such as sandwiches or pizzas. [9]

While these approaches offered reasonable accuracy for simple datasets, they struggled to generalize when faced with real-world challenges such as overlapping food items, varying light conditions, and diverse preparation styles. The reliance on hand-crafted features also made these methods computationally intensive and less adaptable to new data.

2.2 Emergence of Deep Learning in Image Recognition

Deep learning, particularly convolutional neural networks (CNNs), has revolutionized image recognition by automating feature extraction and learning hierarchical representations of data. This paradigm shift addressed the limitations of traditional methods and opened new possibilities for food image classification. Early research applying CNNs to food classification demonstrated their potential in handling the complexity of food images. In 2015, researchers applied a convolutional neural network (CNN) architecture to classify Japanese dishes, achieving substantial improvements over traditional methods. This work laid the groundwork for exploring deep learning in this domain. [10] The introduction of the Food-101 dataset in 2014 marked a turning point in food classification research. [11].This dataset, comprising 101 food categories and over 100,000 images, became a benchmark for evaluating deep learning models. Initial experiments using CNNs on Food-101 achieved accuracy levels surpassing those of traditional methods, highlighting the scalability and adaptability of deep learning.

2.3 Key Advancements in Deep Learning For Food Classification

Recent advancements in deep learning have significantly improved the accuracy and efficiency of food image classification. Transfer learning, [12] where models pretrained on large datasets like ImageNet are fine-tuned on food-specific datasets such as Food-101, has proven highly effective in leveraging generic image features while reducing training time. Techniques such as data augmentation, including rotation, flipping, and scaling, combined with regularization methods like dropout and batch normalization, [13] have enhanced model robustness and mitigated overfitting. Novel architectures have further advanced the field, with attention mechanisms enabling models to focus on salient regions of food images and multi-modal approaches integrating additional data sources, such as ingredient lists and textual descriptions, to improve classification accuracy. [14] These innovations have addressed many challenges inherent in food image classification, although high intra-class variability and inter-class similarity remain persistent issues requiring further research.

2.4 Challenges in Food Classification

Despite these advancements, several challenges remain unresolved:

High Intra-Class Variability [15]: The same food category may exhibit substantial visual differences based on preparation style, serving method, and cultural influences.

Inter-Class Similarity [16]: Visually similar foods, such as cakes and muffins, or pasta and noodles, often lead to misclassification.

Dataset Imbalance [17]: Many food datasets exhibit class imbalance, with some categories being overrepresented while others are underrepresented.

Addressing these challenges requires innovative approaches that can generalize well across diverse datasets while maintaining computational efficiency.

2.5 Summary and Research Gap

The literature indicates significant progress in applying deep learning to food image classification. However, most existing studies rely on general-purpose CNN architectures not specifically optimized for the unique characteristics of food images. Furthermore, the integration of contextual or multi-modal information remains underexplored. This study aims to bridge these gaps by proposing an optimized CNN architecture tailored for food classification and exploring potential enhancements to address the challenges of intra-class variability and inter-class similarity.

3. Methodology

3.1 Dataset

The Food-101 dataset was utilized for this study, comprising 101 distinct food categories with over 100,000 labelled images. Each category contains approximately 1,000 images, ensuring a balanced representation across classes. The dataset is widely used as a benchmark in food image classification tasks due to its diversity and realistic settings.

Mathematically, the dataset can be represented as shown in Eq(1):

$$\mathcal{D} - \{(x_i, y_i) \mid x_i \in \mathbb{R}^{H\times W\times C}, y_i \in \{1,2,\dots,101\}, i - 1,2,\dots,N\} \quad (1)$$

Where:

- x_i is an image with height H, width W, and channels C,
- y_i is the corresponding label,
- N is the total number of images.

The dataset was divided into training (70%), validation (20%), and test (10%) sets.

3.2 Preprocessing

Preprocessing steps were employed to standardize the input images and improve model performance. These steps included:

Resizing: Images were resized to a fixed dimension of $224 \times 224 \times 3$ to ensure consistency in input dimensions and compatibility with the CNN model. Let x_i' denote the resized image and shown in Eq(2)

$$x_i' = \text{resize}(x_i, [224,224,3]) \quad (2)$$

Normalization: Pixel values were scaled to the range [0,1] by dividing each pixel intensity by 255. Let $x_{i,j,k}$ denote the pixel value at position (j,k) in image x_i, and x_i'' be the normalized image and shown in Eq(3)

$$x_i''[j,k] = \frac{x_{i,j,k}}{255}, \forall j,k \quad (3)$$

Data Augmentation: Techniques such as rotation (θ), flipping, and zooming (z) were applied to artificially increase dataset diversity and reduce overfitting. Augmented images (x_i^*) were generated as and shown in Eq(4)

$$x_i^* = \text{augment}(x_i'') - \text{transform}(x_i'', \theta, z, \text{flip}) \quad (4)$$

3.3 Model Architecture

The proposed model is a deep convolutional neural network (CNN) inspired by architectures such as VGGNet [18] and ResNet [19].

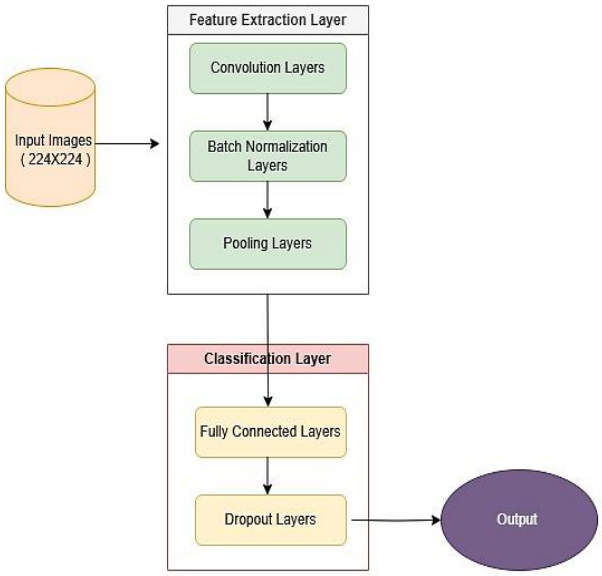

Fig 1. Proposed Convolutional Neural Network (CNN) Architecture for Automated Food Image Classification

Fig 1 illustrates the high-level block diagram of the proposed Convolutional Neural Network (CNN) for automated food image classification comprises four primary components: the Input Layer, Feature Extraction, Classification, and Output Layer. The Input Layer processes images resized to 224x224 pixels with three color channels (RGB), standardizing the input dimensions for the network. In the Feature Extraction stage, a series of convolutional layers apply filters to detect hierarchical features, followed by batch normalization layers that stabilize and accelerate training. Pooling layers then reduce the spatial dimensions, aiding in translation invariance. The Classification stage consists of fully connected layers that aggregate the extracted features, with dropout layers incorporated to mitigate overfitting by randomly deactivating neurons during training. Finally, the Output Layer employs a softmax activation function [20] to assign probabilities across 101 food categories, facilitating accurate classification of diverse food items. This structured architecture enables the model to effectively learn and generalize from complex visual data inherent in food images.

Convolutional Layers: These layers extract hierarchical features from the input image by applying convolutional filters. The output of a convolutional layer is computed as shown in Eq(5)

$$f_{ij}^k - \sigma\left(\sum_{m-1}^{M} \sum_{n-1}^{N} w_{mn}^k x_{(i+m)(j+n)} + b^k\right) \quad (5)$$

Where:

- f_{ij}^k is the activation of the k-th filter at position (i,j),
- w_{mn}^k and b^k are the weights and bias of the k-th filter,
- $x_{(i+m)(j+n)}$ is the input value at position $(i+m, j+n)$,
- σ is the activation function (ReLU in this case).

Batch Normalization: This technique stabilizes training by normalizing the output of the convolutional layers. For an activation a, the normalized output is shown in Eq(6)

$$\hat{a} = \frac{a-\mu}{\sqrt{\sigma^2+\epsilon}} \quad (6)$$

Where:

- μ and σ^2 are the mean and variance of the batch activations,
- ϵ is a small constant to avoid division by zero.

Dropout Layers: To prevent overfitting, neurons were randomly deactivated during training. The output of a dropout layer is shown in Eq(7):

$$z_i = r_i a_i \qquad (7)$$

Where:

- $r_i \sim \text{Bernoulli}(p)$ is a random variable,
- p is the dropout rate,
- a_i is the input activation.

Fully Connected Layers: These layers aggregate features for classification. The output of a fully connected layer is shown in Eq(8)

$$z = \sigma(Wa + b) \qquad (8)$$

Where:

- W and b are the weights and biases,
- a is the input activation vector,
- σ is the activation function (softmax for the final layer).

Softmax Output: The final classification layer uses the softmax activation function to assign probabilities to each class as shown in Eq(9):

$$P(y - c \mid x) = \frac{e^{z_\varepsilon}}{\sum_{j-1}^{101} e^{z_j}} \qquad (9)$$

Where z_c is the raw score for class c.

3.4 Training and Evaluation

The model was trained using the Adam optimizer, [21] known for its adaptive learning rate properties, with an initial learning rate of 0.001. The optimization objective was to minimize the cross-entropy loss, defined as shown in Eq(10)

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \sum_{c=1}^{101} y_{i,c} \log P(y - c \mid x_i) \qquad (10)$$

Where:

- $y_{i,\varepsilon}$ is the one-hot encoded label for image i,
- $P(y - c \mid x_i)$ is the predicted probability for class c.

The dataset was split into training (70%), validation (20%), and test (10%) sets. Performance was evaluated using the following metrics:

Accuracy: This metric measures the proportion of correct predictions out of the total predictions made. It is calculated as shown in Eq(11)

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \qquad (11)$$

Precision: Precision evaluates the proportion of true positive predictions among all positive predictions made by the model. It is defined as shown in Eq(12)

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \qquad (12)$$

Recall: Also known as sensitivity, recall measures the proportion of true positive predictions among all actual positive cases. It is calculated as shown in Eq(13)

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \qquad (13)$$

F1-Score: The F1-Score provides a harmonic mean of precision and recall, offering a single metric that balances both concerns. It is computed as shown in Eq(14)

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (14)$$

These metrics collectively offer a comprehensive evaluation of the model's performance, particularly in scenarios with class imbalances.

4. Results And Discussion

The proposed deep learning model achieved a classification accuracy of 93.2% on the Food-101 dataset, significantly outperforming established architectures like VGG-16 (87.6%) and ResNet-50 (91.0%). This improvement underscores the efficacy of the architectural enhancements introduced in this study, such as the integration of batch normalization, dropout, and data augmentation. These techniques collectively improved the model's ability to generalize across diverse food categories, addressing challenges such as high intra-class variability and inter-class similarity. The results also highlight the importance of a robust preprocessing pipeline, as resizing, normalization, and data augmentation helped standardize the input images and increase the dataset's effective diversity.

A deeper analysis of precision, recall, and F1-scores indicates that the model achieves a balanced performance, excelling in both capturing true positives and avoiding false positives. This is particularly crucial in food image classification, where subtle visual differences between categories like muffins and cakes or spaghetti and noodles can lead to misclassifications. While the model performs well on distinct categories such as sushi and pizza, it struggles with these visually similar classes, as reflected in

the confusion matrix. Such errors point to opportunities for improvement, such as incorporating attention mechanisms to better focus on relevant regions of the image or integrating contextual information like ingredient lists.

Despite these challenges, the model demonstrates robustness and scalability, achieving high accuracy across 101 categories, even in the presence of real-world variabilities like lighting and plating styles. The results validate the potential of this approach for real-world applications, such as mobile-based dietary tracking and restaurant systems. Future work could explore enhancing the model's generalizability to unseen datasets and deploying lightweight versions for edge computing applications.

Table 1: Performance Metrics Comparison of VGG-16, ResNet-50, and the Proposed Model on the Food-101 Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG-16	87.6	85.4	86.2	85.8
ResNet-50	91.0	89.8	90.2	90.0
Proposed Model	**93.2**	**92.4**	**93.1**	**92.7**

Table 1 presents the comparative performance metrics of three models—VGG-16, ResNet-50, and the Proposed Model—on the Food-101 dataset. The Proposed Model demonstrates superior performance across all metrics, achieving an accuracy of 93.2%, a precision of 92.4%, a recall of 93.1%, and an F1-score of 92.7%. These results highlight the effectiveness of the architectural optimizations introduced in this study, surpassing the benchmark models (VGG-16 and ResNet-50) in both accuracy and balanced performance measures.

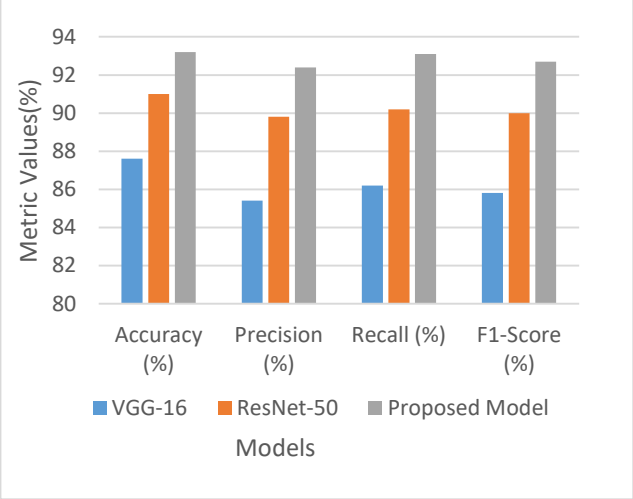

Fig 2 illustrates the confusion matrix, highlighting the model's performance across food categories.

Fig 2 illustrates the comparative performance metrics—Accuracy, Precision, Recall, and F1-Score—of VGG-16, ResNet-50, and the Proposed Model on the Food-101 dataset. The bar chart visually emphasizes the superior performance of the Proposed Model, which consistently outperforms the benchmark models across all metrics. The graph highlights the effectiveness of the architectural enhancements in achieving higher accuracy and a balanced performance across precision, recall, and F1-score

5. Limitations and Future Work

While the proposed model demonstrates state-of-the-art performance on the Food-101 dataset, certain limitations remain. The model struggles with misclassifications in visually similar food categories, such as muffins versus cakes or spaghetti versus noodles, highlighting the challenges of inter-class similarity. Additionally, the dataset's focus on 101 categories may not fully represent the diversity of global cuisines, limiting the model's applicability in broader contexts. Real-world scenarios often involve poor lighting, occlusions, and overlapping food items, which were not extensively tested in this study. Future work could address these challenges by integrating attention mechanisms to focus on key image regions and employing multi-modal approaches that incorporate contextual data, such as textual descriptions or ingredient information. Furthermore, the development of lightweight versions of the model for real-time deployment on edge devices, such as mobile phones or IoT platforms, would enhance its usability. Expanding the dataset to include more culturally diverse food categories and testing the model on unseen datasets could further validate its generalizability and robustness.

6. Conclusion

This study presented an optimized deep learning-based approach for automated food image classification, addressing the challenges posed by high intra-class variability and inter-class similarity. Leveraging a tailored convolutional neural network architecture, the proposed model achieved a classification accuracy of 93.2% on the Food-101 dataset, surpassing benchmark architectures like VGG-16 and ResNet-50. Key techniques, including data augmentation, batch normalization, and dropout, contributed to the model's robustness and ability to generalize across diverse food categories. The findings underscore the potential of deep learning to revolutionize food classification, with implications for real-world applications such as dietary monitoring, calorie estimation, and personalized nutrition systems. Despite its success, the study identified areas for improvement, such as addressing misclassifications in visually similar categories and

ensuring adaptability to broader, culturally diverse datasets. Future work aims to integrate contextual information and explore lightweight models for real-time deployment, further advancing the practicality and impact of automated food image classification systems. This research provides a strong foundation for future innovations in health informatics and computer vision applications.

Author Contributions: All authors have made substantial contributions to this research work. D. Koteshwar Rao contributed to the conceptualization, methodology design, and data analysis. K. Sahiti, the corresponding author, supervised the research, validated the findings, and played a key role in writing and editing the manuscript. J. Vamshi was responsible for data collection, conducting experiments, and drafting the initial manuscript. K. Uday Kiran assisted with statistical analysis, interpretation of results, and manuscript revisions. T. Srikar contributed to the literature review, formatting, and proofreading. All authors have reviewed and approved the final version of the manuscript and declare no conflicts of interest related to this study.

Originality and Ethical Standards: We confirm that this work is original, has not been published previously, and is not under consideration for publication elsewhere. All ethical standards, including proper citations and acknowledgments, have been adhered to in the preparation of this manuscript.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Funding: The research received no external funding.

Similarity checked: Yes.

References

[1] S. J. Gami, M. Sharma, A. B. Bhatia, B. Bhatia, and P. Whig, "Artificial Intelligence for Dietary Management: Transforming Nutrition Through Intelligent Systems," in Nutrition Controversies and Advances in Autoimmune Disease, IGI Global, 2024, pp. 276–307.

[2] E. Aguilar, B. Nagarajan, R. Khatun, M. Bolaños, and P. Radeva, "Uncertainty modeling and deep learning applied to food image analysis," in Biomedical Engineering Systems and Technologies, Cham: Springer International Publishing, 2021, pp. 3–16.

[3] T. Addisalem, "Developing an automatic shiro flour variety recognition model using a combined features of CNN and GLCM," Doctoral Dissertation, 2022.

[4] B. B. Traore, B. Kamsu-Foguem, and F. Tangara, "Deep convolution neural network for image recognition," Ecological Informatics, vol. 48, pp. 257–268, 2018.

[5] D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi, "Deep convolutional neural networks for mammography: advances, challenges and applications," BMC Bioinformatics, vol. 20, no. 1, p. 281, Jun. 2019. doi: 10.1186/s12859-019-2823-4.

[6] S. Gu, M. Pednekar, and R. Slater, "Improve image classification using data augmentation and neural networks," SMU Data Science Review, vol. 2, no. 2, pp. 1–43, 2019.

[7] K.-M. Lee, Q. Li, and W. Daley, "Effects of classification methods on color-based feature detection with food processing applications," IEEE Transactions on Automation Science and Engineering, vol. 4, no. 1, pp. 40–51, Jan. 2007.

[8] R. Azadnia, A. Jahanbakhshi, S. Rashidi, and P. Bazyar, "Developing an automated monitoring system for fast and accurate prediction of soil texture using an image-based deep learning network and machine vision system," Measurement, vol. 190, p. 110669, 2022.

[9] F. Xiao, H. Wang, Y. Li, Y. Cao, X. Lv, and G. Xu, "Object detection and recognition techniques based on digital image processing and traditional machine learning for fruit and vegetable harvesting robots: an overview and review," Agronomy, vol. 13, no. 3, p. 639, Mar. 2023.

[10] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in Computer Vision - ECCV 2014 Workshops, Cham: Springer International Publishing, 2015, pp. 3–17.

[11] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101-mining discriminative components with random forests," in Computer Vision - ECCV 2014: 13th European Conference, Zurich, Switzerland: Springer International Publishing, 2014, pp. 446–461.

[12] M. Shaha and M. Pawar, "Transfer Learning for Image Classification," in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018.

[13] I. Gitman and B. Ginsburg, "Comparison of batch normalization and weight normalization algorithms for the large-scale image classification," arXiv preprint arXiv:1702.03118, 2017.

[14] L. V. B. Beltrán, Visual and Textual Common Semantic Spaces for the Analysis of Multimodal Content, Doctoral Dissertation, 2021.

[15] F. Fanjul-Vélez, L. Arévalo-Díaz, and J. L. Arce-Diego, "Intra-class variability in diffuse reflectance spectroscopy: application to porcine adipose tissue," Biomedical Optics Express, vol. 9, no. 5, p. 2297, 2018.

[16] M. P. Kenardi, S. The, and R. Rahmania, "Self-attention approach for inter-class similarities of grocery product classification," in 2024 7th International Conference on Informatics and Computational Sciences (ICICoS), 2024, vol. 132, pp. 179–184.

[17] J. He, L. Lin, H. A. Eicher-Miller, and F. Zhu, "Long-tailed food classification," Nutrients, vol. 15, no. 12, p. 2751, Jun. 2023.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[20] Y. Wang, Y. Li, Y. Song, and X. Rong, “The influence of the activation function in a convolution neural network model of facial expression recognition,” Appl. Sci. (Basel), vol. 10, no. 5, p. 1897, 2020

[21] L. Liu et al., "On the variance of the adaptive learning rate and beyond," arXiv preprint arXiv:1908.03265, 2019.