Cureus Journal of Computer Science

Open Access Original Article

EfficientNetB0-CatBoost: Deep Learning With Categorical Boosting for Food Image Recognition

Pranjal Kumar Singh [1] , Seba Susan [1]

1. Information Technology, Delhi Technological University, Delhi, IND

Corresponding author: Seba Susan, seba_406@yahoo.in

Received 03/15/2025
Review began 03/18/2025
Review ended 10/16/2025
Published 10/21/2025

© Copyright 2025
Singh et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI:
https://doi.org/10.7759/s44389-025-03791-2

Abstract

Research into machine learning and deep learning methods for image categorization has been sparked by the explosion of visual content available on digital platforms, with particular interest in domains such as food image recognition. With so many food-related images available on the internet and social media platforms, automated food image categorization is a major difficulty. Deep Convolutional Neural Network architectures offer a promising approach by leveraging pre-trained models and transfer learning to extract meaningful image representations. EfficientNetB0 is an example of a pre-trained deep learning architecture that has demonstrated impressive ability in extracting high-level features from images. This study explores the effectiveness of using deep features extracted from the average pooling layer of a fine-tuned EfficientNetB0 model for the classification of complex food images using the Food-101 dataset. Subsequently, the features extracted are passed as input to a Categorical Boosting (CatBoost) classifier. To assess performance, we compare the classification accuracies of CatBoost, Support Vector Machines with various kernels, and a Random Forest classifier, all utilizing the same set of extracted features, against the accuracy of the end-to-end fine-tuned EfficientNetB0 model. Our experiments reveal that the proposed EfficientNetB0-CatBoost framework achieves the highest classification accuracy for classifying complex food images, highlighting the potential of CatBoost for efficient classification of deep features.

Categories: AI applications, Image Processing and Analysis, Deep Learning
Keywords: food image classification, deep learning, categorical boosting, machine learning, efficientnet

Introduction

The integration of deep learning and machine learning methodologies has ushered in a new era of problem-solving across various fields [1]. Machine learning techniques provide powerful tools for detecting complex patterns in large datasets. By utilizing these techniques, it is possible to identify minute variations in colors, textures, and forms, enabling highly accurate classification of food images [2]. On the other hand, deep learning has emerged as a powerful approach for handling high-dimensional and complex data, demonstrating unparalleled capability in modeling intricate patterns and relationships within complex datasets [3]. When it comes to feature extraction, convolutional neural networks (CNNs) have shown themselves to be highly proficient, collecting hierarchical representations of images with ease [4]. The collaboration between deep neural networks and machine learning algorithms has significantly expedited and enhanced the accuracy of classification processes, marking a notable advancement in addressing complex challenges within diverse domains [5]. The classification of deep features extracted from transfer-learned pre-trained networks has yielded promising results in the past [6-7]. This integration has brought about transformative changes, particularly in the realm of food image classification [8], a pivotal aspect in today's digitally driven landscape characterized by the widespread sharing of visual content.

There are several researchers who have applied deep learning and/or machine learning for classification of food-related images. We discuss some of these works next. Medus et al. [9] presented a method for using CNNs in heat-sealed food trays for real-time packaging control using hyperspectral imaging systems. The CNN architecture and training methods achieve a global fault detection accuracy of over 94%. The algorithm can be combined with automated production line control at a maximum speed of 14 food trays per second, resulting in a total computation time of 70-105 ms. Bossard et al. [10] introduced a unique application of Random Forest (RF) for automated food image recognition. The process shares knowledge across all classes by simultaneously mining discriminative portions. To increase efficiency, the authors utilized patches that are aligned with image superpixels as components. The model achieved 50.76% accuracy, outperforming all compared methods except CNNs, including Support Vector Machine (SVMs) on Improved Fisher Vectors and existing part-mining algorithms. Jiang et al. [11] provided a three-step algorithm that combines nutritional evaluation and food recognition using deep models. The system classifies food items into categories, analyzes nutritional elements, and identifies multi-item food items using a three-step algorithm. This results in clear insights into good eating habits and daily recipe guidance for users. Pandey et al. [12] presented a multilayered CNN pipeline-based automated food identification system. Preprocessed images are utilized to train and fine-tune the network, and for increased accuracy, the filter outputs are fused. According to experimental results, the proposed ensemble architecture outperformed state-of-the-art models on a large real-world food identification database and a new collection of Indian cuisine images. Aguilar et al. [13] proposed a strategy to improve food image recognition performance by combining multiple classifiers based on convolutional models. Using the maximum similarity between decision templates and test image decision profiles as the similarity measure, the method outperforms standalone CNN models. A five-layer CNN was utilized by early researchers [14] to identify 5,822 images from 10 food categories in a small-scale food image dataset. When it came to classification, the CNN model outperformed the bag-of-features model and SVM, with an overall accuracy of 74%. The accuracy was increased to more than 90% by employing geometric modifications in data augmentation approaches. Because there were not enough training data, the CNN model had trouble fitting. Rather than simply increasing the number of training epochs, larger improvements may be achieved by collecting more training data and optimizing the network architecture and hyperparameters.

Deep pre-trained CNNs harness the power of transfer learning for adaptation to smaller classification tasks [15]. These networks are pre-trained on a larger dataset – ImageNet – that contains a million images pertaining to different object categories. Transfer learning in the current application would involve the fine-tuning of ImageNet weights on the smaller food image datasets, facilitating the transfer of knowledge from one domain to another. We cite some examples of deep pre-trained CNNs that have been successfully adapted for food image recognition. Some of the earliest works simply focused on distinguishing between food and non-food images; Singla et al. [16] employed the GoogLeNet pre-trained model to achieve the task, with high classification accuracies being obtained. It becomes especially challenging to differentiate food items when their colors and textures appear similar in images. Due to high inter-class similarity and substantial intra-class variation, food image classification remains a particularly difficult challenge in computer vision. In a recent work [17], the authors compared five pre-trained networks, InceptionV3, DenseNet121, MobileNet, EfficientNetB0, and Xception, and found that EfficientNetB0, InceptionV3, and Xception performed the best for classification of images from the Food-101 dataset. DenseNet121 gave the worst performance as compared to other models. Konstantakopoulos et al. [18] successfully classified Mediterranean food images using the EfficientNet pre-trained model. They achieved 83.4% accuracy by fine-tuning the EfficientNet model and 78.9% accuracy without fine-tuning. A similar experiment was conducted by Liu [19] who compared EfficientNetB0 and ResNet50 pre-trained models on images from the Food-101 dataset. The classification accuracy of EfficientNetB0 (75.14%) was found significantly higher than ResNet50 (62.34%). Ittisoponpisan et al. [20] found EfficientNet to be effective for the classification of Thai food images.

This study specifically endeavors to assess the classification accuracy achieved by categorical boosting (CatBoost) (machine learning) utilizing deep features extracted from the EfficientNetB0 model (deep learning) for food image recognition from a complex benchmark dataset.

Materials And Methods

A scrutiny of literature reveals that several experiments on food image classification have been conducted on the deep pre-trained EfficientNet model, with high accuracies (>80%) being achieved on complex images from the benchmark Food-101 dataset. The EfficientNetB0 model [21] demonstrates an advanced architecture intended to effectively extract complex patterns from input images. The input flow consists of convolutional layers, the middle flow consists of repeated mobile inverted MBConv blocks [22-23], and the exit flow consists of a convolutional, global average pooling and fully connected layers. For deep feature extraction, we use the Global Average Pooling layer, which is the fourth penultimate layer of the fine-tuned EfficientNetB0 model. Figure *1* illustrates the process of extracting feature maps from this model.

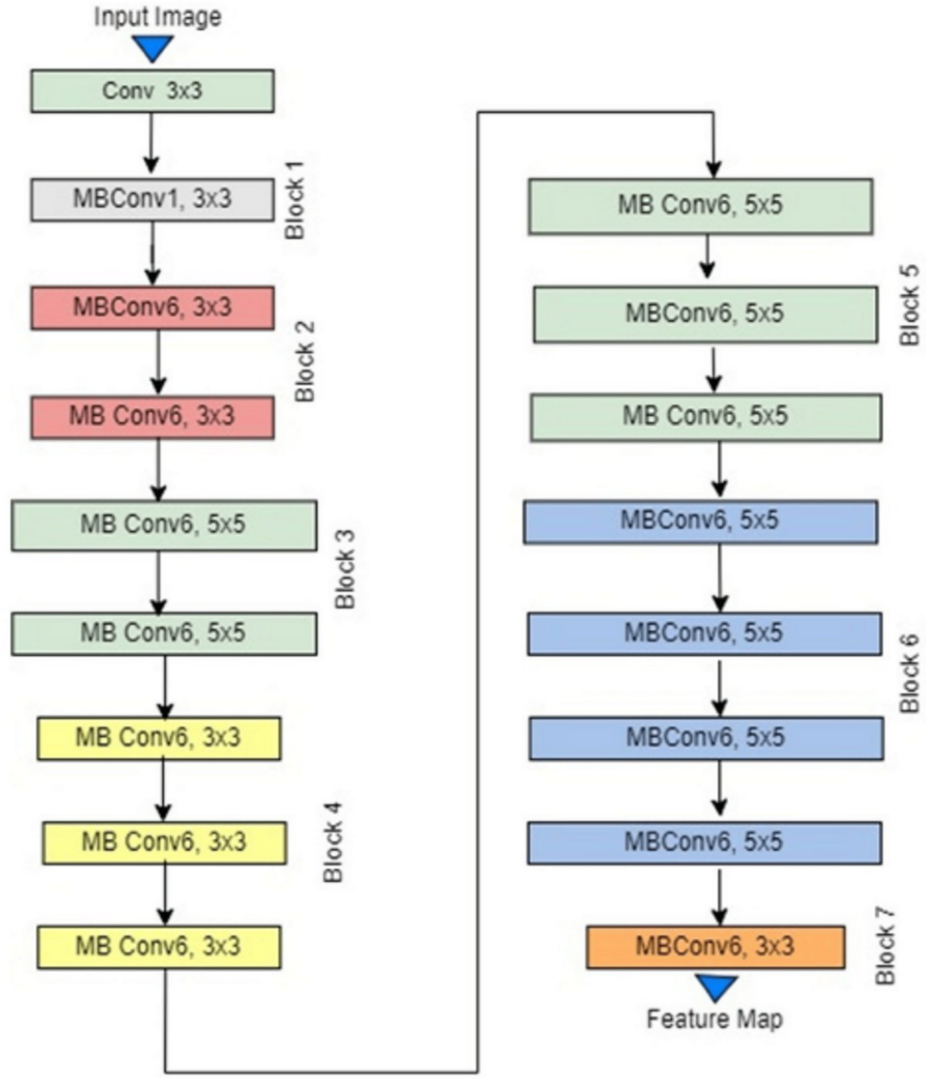

FIGURE 1: The deep pre-trained EfficientNetB0 model fine-tuned for the food image classification task.

Adapted from Iqbal et al. [24].

The EfficientNetB0 model is especially well suited for food image recognition due to the following reasons:

1. EfficientNetB0 is well suited for food image recognition due to its compound scaling, which balances network depth, width, and input resolution for optimal accuracy and efficiency. Compound scaling increases input resolution, allowing the model to preserve and analyze more pixel-level information, which is essential for distinguishing visually similar dishes.

2. EfficientNet's balanced scaling helps it learn both fine-grained details and high-level patterns more effectively than ResNet (which scales depth) or DenseNet (which primarily increases connectivity).

3. Its use of MBConv blocks and squeeze-and-excitation mechanisms enables effective extraction of fine-grained visual features – crucial for distinguishing similar food items.

4. The Global Average Pooling layer captures a high-level representation that captures global characteristics of the food item – such as overall shape, color distribution, and texture patterns, while reducing overfitting.

5. Its lightweight architecture offers high performance with fewer parameters, making it ideal for practical applications. Despite its compact size, it maintains high classification accuracy on food datasets such as

Food-101, where images have complex visual cues and subtle class differences.

Overall, EfficientNetB0 achieves strong accuracy on food datasets while remaining computationally efficient. In the current work, we aim to enhance the deep learning performance of the EfficientNet model for food image recognition by using it as a feature extractor and leveraging the effectiveness of categorical boosting to classify the extracted deep features. Previous works have highlighted the significance of the top-most average pooling layer in deep pre-trained networks as an efficient feature extractor for various computer vision applications [25-26]. EfficientNetB0 has 237 layers and is known to be a very deep model, which is also a reason for the high recognition accuracies it achieves. In the current experiments, the top-56 layers of EfficientNetB0 are fine-tuned on the target food image dataset, while the lower 181 layers retain the original ImageNet weights. The pipeline of the proposed methodology for food image recognition is illustrated in Figure 2.

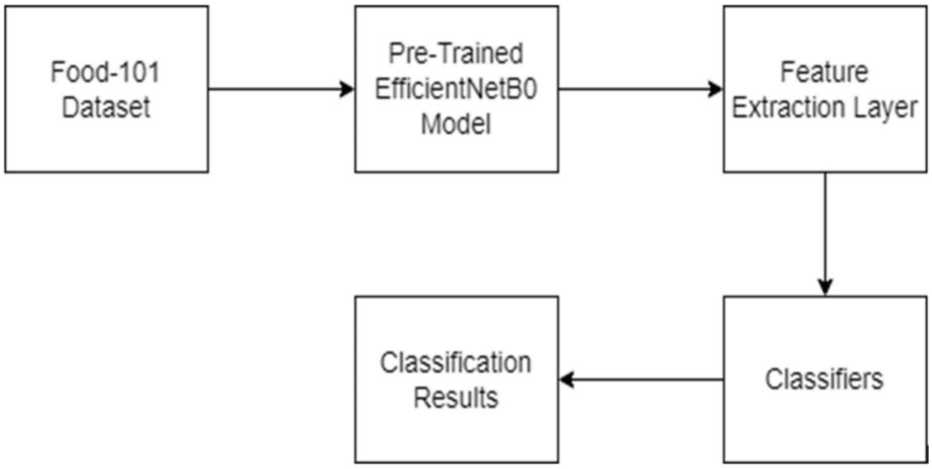

FIGURE 2: Pipeline for the proposed EfficientNet-CatBoost framework.

CatBoost, Categorical Boosting

The pre-trained EfficientNetB0 model is harnessed to glean intricate features from the Food-101 dataset images. These features are derived from the Global Average Pooling layer, resulting in a feature vector initially shaped as (3, 3, 1280). To prepare these features for downstream analysis, we flatten the vector, transforming it into a tensor of shape 11,520. Each image in the dataset undergoes this transformation, generating a set of 11,520 features. The activations of the global average pooling payer of the EfficientNetB0 model are subsequently passed as input to the classification phase that comprises of a CatBoost classifier. Yandex researchers created the robust gradient boosting library CatBoost [27] to effectively handle categorical information in machine learning applications. It is notable for being scalable, resilient, and capable of achieving excellent results on a variety of datasets. CatBoost is a flexible and effective gradient boosting toolkit that performs well with categorical data, reduces overfitting, and generates reliable and comprehensible models. To test the efficacy of the proposed CatBoost model on the extracted features, we compare its performance with other machine learning classifiers, such as SVM and RF that are reported to have achieved acceptable accuracies for food image classification when trained on popular image features such as Scale Invariant Feature Transform, Speeded Up Robust Features, and Histogram of Oriented Gradients [28].

Results

The following food categories from the Food-101 dataset [10] were used for the experimentation: Apple Pie, Breakfast Burrito, Carrot Cake, Cheesecake, Chicken Wings, Chocolate Cake, Donuts, French Fries, French Toast, Fried Rice, Garlic Bread, Hot Dog, Ice Cream, Lasagna, Macaroni & Cheese, Omelette, Onion Rings, Pancakes, Pizza, Ramen, Red Velvet Cake, Samosa, Sushi, Tacos, and Waffles, twenty five in all. Each class comprised 750 training samples and 250 test samples. Therefore, there were 18,750 training images and 6,250 test images available for the training and testing phases, respectively. The dataset displayed high inter-class similarity and large intra-class variation that may possibly confuse the machine learning and deep learning models. Some samples are displayed in Figure 3, which illustrate the subtle similarities in colors and textures of certain food categories such as Garlic Bread, Pizza, and Ramen.

FIGURE 3: Some samples from the Food-101 dataset showcasing the complexity of the food images.

All experiments are performed in Python (3.7 version). The images were resized to 224×224 (RGB color) prior to application to EfficientNetB0. The EfficientNetB0 model was trained on the Food-101 dataset for around 40 hours using an Intel Core i7 11th generation CPU with 16GB RAM, without GPU acceleration. The hyperparameter selection process is crucial to the performance of both deep learning models and traditional classifiers. To ensure robustness and generalizability, we employed a systematic tuning approach using grid search, which allowed us to identify optimal configurations, rather than relying on arbitrary choices. For the EfficientNetB0 model, training was performed using the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.1 and momentum of 0.9. A dropout rate of 0.5 was employed to mitigate overfitting and promote more robust learning. The model was trained with a batch size of 32 across multiple epochs, allowing for iterative refinement of its weights. Additionally, a flexible learning rate schedule was maintained during training to adaptively optimize performance across epochs. For the CatBoost classifier, key parameters include the number of trees, tree depth, and learning rate-all of which significantly influence model complexity and convergence behavior. Specifically, we set the tree depth to 5, optimized the learning rate to 0.01, and used 1,000 trees. To prevent overfitting and ensure robustness, we applied early stopping and L2 regularization, which collectively improved the classifier's accuracy and stability on food image data. For SVM, hyperparameters were tuned using grid search, resulting in gamma = 0.1 and C = 1 as the optimal configuration. In the case of the RF classifier, we configured the model with 200 trees (number of estimators) and a maximum depth of 30 per tree, balancing performance with computational efficiency.

Table *1* presents the accuracy results of various approaches explored in our study. We compare with the baseline EfficientNetB0 model that has proved to outperform other deep learning models like ResNet, DenseNet and MobileNet for food image classification [17-20]. For the classification phase, we test the efficacy of CatBoost, SVM (linear, Gaussian, and polynomial kernels), and RF classifiers on the extracted deep features.

Model	Methodology	Accuracy
(Baseline) EfficientNetB0	Deep learning	81.49%
EfficientNetB0-Random Forest	Deep learning + Machine learning	74.15%
EfficientNetB0-SVM (linear)	Deep learning + Machine learning	84.61%
EfficientNetB0-SVM (Gaussian)	Deep learning + Machine learning	82.46%
EfficientNetB0 (Polynomial)	Deep learning + Machine learning	79.45%
EfficientNetB0-CatBoost (Proposed)	Deep learning + Machine learning	87.53%

TABLE 1: Comparison of classification performance across different deep learning models on the Food-101 food image recognition task.

Discussion

Figure *4* presents the training accuracy and loss curves of the proposed model, illustrating the learning progression over epochs. The accuracy curve reflects how well the model improves in correctly classifying training samples, while the loss curve indicates the model's error reduction during training. A smooth decline in loss and a steady rise in accuracy typically signify effective learning and convergence.

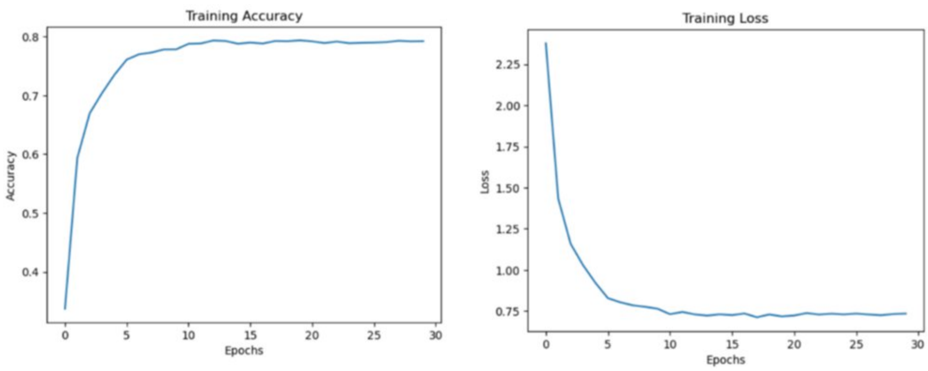

FIGURE 4: Training accuracy and loss curves of the proposed model, illustrating the learning dynamics and convergence behavior over successive epochs.

As observed from Table *1*, the baseline EfficientNetB0 model achieved an accuracy of 81.49%, serving as a reference point for hybrid approaches. Among the machine learning classifiers, CatBoost significantly outperformed others, reaching the highest accuracy of 87.53%. SVM with a linear kernel also showed strong performance at 84.61%, surpassing the baseline. SVM with a Gaussian kernel and Polynomial kernel achieved moderate results at 82.46% and 79.45%, respectively. The RF model lagged behind with only 74.15%, indicating a poor fit for the extracted deep features. Among all methods, the proposed EfficientNet-CatBoost framework achieved the highest accuracy of 87.53%, demonstrating its superior performance for food image classification. Overall, combining deep learning with advanced machine learning methods, particularly CatBoost, led to improved classification performance.

Conclusions

This study explores the synergistic combination of strong classification algorithms, such as CatBoost with deep feature extraction from very deep pre-trained models. Our goal is to improve performance and accuracy in food image classification tasks by determining the best feature extraction and classification method combination via careful experimentation and review. Our research concludes by demonstrating how well the deep learning models – specifically, the EfficientNetB0 architecture – perform for challenges involving the categorization of food images. The EfficientNetB0 architecture is trained on 18,750 images of the benchmark Food-101 dataset encompassing twenty five food categories. We sought to assess the classification accuracy by extracting deep features from the global average pooling layer of the EfficientNetB0 model, and passing them into conventional machine learning classifiers like CatBoost, SVM, and RF, for classification. Our findings show that CatBoost outperforms SVM and RF, achieving the highest test accuracy of 87.53%, demonstrating its effectiveness in classifying food images using deep extracted features. This suggests that combining categorical boosting with deep learning can enhance classification performance. However, to ensure broader applicability, future work should explore potential dataset biases, overfitting risks, and the scalability of this framework in real-world food recognition systems.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Seba Susan, Pranjal Kumar Singh

Acquisition, analysis, or interpretation of data: Seba Susan, Pranjal Kumar Singh

Drafting of the manuscript: Seba Susan, Pranjal Kumar Singh

Critical review of the manuscript for important intellectual content: Seba Susan, Pranjal Kumar Singh

Supervision: Seba Susan

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Liu R, Pan D, Xu Y, et al.: A deep learning-machine learning fusion approach for the classification of benign, malignant, and intermediate bone tumors. European Radiology. 2022, 32:1371-1383. 10.1007/s00330-021-08195-z
2. Sharifani K, Amini M: Machine learning and deep learning: A review of methods and applications. World Information Technology and Engineering Journal. 2023, 10:3897-3904.
3. Agarwal A, Susan S: Attention-augmented squeeze-and-excitation enhanced mobile network for occluded facial expression recognition in resource-constrained environments. Signal, Image and Video Processing. 2025, 19:687. 10.1007/s11760-025-04278-2
4. Dara S, Tumma P: Feature extraction by using deep learning: A survey. 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India. 2018, 1795-1801. 10.1109/ICECA.2018.8474912
5. Saini M, Susan S: Bag-of-Visual-Words codebook generation using deep features for effective classification of imbalanced multi-class image datasets. Multimedia Tools and Applications. 2021, 80:20821-20847. 10.1007/s11042-021-10612-w
6. Kaushik A, Susan S: Metric learning with deep features for highly imbalanced face dataset. International Conference on Innovative Computing and Communications. Khanna A, Gupta D, Bhattacharyya S, Hassanien AE, Anand S, Jaiswal A (ed): Springer, Singapore; 2022. 3:639-646. 10.1007/978-981-16-3071-2_52
7. Agarwal A, Susan S: Two-way transfer learning using mobile-dense network for occluded face emotion recognition: a soft decision fusion approach. Pattern Recognition. Palaiahnakote S, Schuckers S, Ogier JM, Bhattacharya P, Pal U, Bhattacharya S (ed): Springer, Cham; 2024. 15616:63-71. 10.1007/978-3-031-87663-9_5
8. Wei P, Wang B: Food image classification and image retrieval based on visual features and machine learning. Multimedia Systems. 2022, 28:2053-2064. 10.1007/s00530-020-00673-6
9. Medus LD, Saban M, Francés-Víllora JV, Bataller-Mompeán M, Rosado-Muñoz A: Hyperspectral image classification using CNN: Application to industrial food packaging. Food Control. 2021, 125:107962. 10.1016/j.foodcont.2021.107962
10. Bossard L, Guillaumin M, Van Gool L: Food-101-mining discriminative components with random forests. Computer Vision - ECCV 2014. Fleet D, Pajdla T, Schiele B, Tuytelaars T (ed): Springer, Cham; 2014. 8694:446-461. 10.1007/978-3-319-10599-4_29
11. Jiang L, Qiu B, Liu X, et al.: DeepFood: food image analysis and dietary assessment via deep model. IEEE Access. 2020, 8:47477-47489. 10.1109/access.2020.2973625
12. Pandey P, Deepthi A, Mandal B, Puhan NB: FoodNet: Recognizing foods using ensemble of deep networks. IEEE Signal Processing Letters. 2017, 24:1758-1762. 10.1109/lsp.2017.2758862
13. Aguilar E, Bolaños M, Radeva P: Food recognition using fusion of classifiers based on CNNs. Image Analysis and Processing - ICIAP 2017. Battiato S, Gallo G, Schettini R, Stanco F (ed): Springer, Cham; 2017. 10485:213-224. 10.1007/978-3-319-68548-9_20

14. Min NM, Khaing MM: Classification for Myanmar food image using DCNN. 2024 IEEE Conference on Computer Applications (ICCA). 2024, 1-6. 10.1109/ICCA62361.2024.10532946
15. Sethi D, Arora K, Susan S: Transfer learning by deep tuning of pre-trained networks for pulmonary nodule detection. 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS). 2020, 168-173. 10.1109/ICIIS51140.2020.9342686
16. Singla A, Yuan L, Ebrahimi T: Food/non-food image classification and food categorization using pre-trained googlenet model. MADiMa '16: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management. 2016, 3-11. 10.1145/2986035.2986039
17. Singh PK, Susan S: Transfer learning using very deep pre-trained models for food image classification. 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India. 2023, 1-6. 10.1109/ICCCNT56998.2023.10307479
18. Konstantakopoulos FS, Georga EI, Fotiadis DI: Mediterranean food image recognition using deep convolutional networks. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico. 2021, 1740-1743. 10.1109/EMBC46164.2021.9630481
19. Liu Y: Automatic food recognition based on efficientnet and ResNet. Journal of Physics: Conference Series. 2023, 2646:012037. 10.1088/1742-6596/2646/1/012037
20. Ittisoponpisan S, Kaipan C, Ruang-on S, Thaiphan R, Songsri-in K: Pushing the accuracy of Thai food image classification with transfer learning. Engineering Journal. 2022, 26:57-71. 10.4186/ej.2022.26.10.57
21. Tan M, Le QV: EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning. 2019, 97:6105-6114.
22. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC: MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018, 4510-4520.
23. Tan M, Chen B, Pang R, Vasudevan V, Sandler M, Howard A, Le QV: MnasNet: Platform-aware neural architecture search for mobile. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019, 2820-2828.
24. Iqbal A, Jaffar MA, Jahangir R: Enhancing brain tumour multi-classification using Efficient-Net B0-based intelligent diagnosis for Internet of Medical Things (IoMT) applications. Information. 2024, 15:489. 10.3390/info15080489
25. Bodapati JD, Veeranjaneyulu N: Feature extraction and classification using deep convolutional neural networks. Journal of Cyber Security and Mobility. 2018, 8:261-276. 10.13052/2245-1439.825
26. Khan MA, Kadry S, Zhang YD, Akram T, Sharif M, Rehman A, Saba T: Prediction of COVID-19 - pneumonia based on selected deep features and one class kernel extreme learning machine. Computers & Electrical Engineering. 2021, 90:106960. 10.1016/j.compeleceng.2020.106960
27. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A: CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems. 2018, 31:6638-6648.
28. Manzoor MA, Morgan Y: Vehicle make and model recognition using random forest classification for intelligent transportation systems. 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA. 2018, 148-154. 10.1109/CCWC.2018.8301714