

Comparative Analysis of State-of-the-Art Deep Learning Models for Food Image Recognition

Atheer Mohammed Al-Shahrani

College of Manag Collage of Informatics, Midocean University, Union of Comoros.

Abstract

Image recognition with food has emerged as a significant field of research in computer vision because of the increasing number of applications in the healthcare, nutrition monitoring, dietary assessment and smart food systems. Nevertheless, proper categorization of food images is a challenge because there is a high level of intra-class variance and visual similarity among various food categories.

The paper is a comparative analysis of the deep learning state-of-the-art models in food image recognition. Three popular convolutional neural network structures, namely ResNet50, MobileNetV2, and EfficientNetB0, were tested on three publicly available datasets with the complexity of the classification task, namely Food-11, Food-34, and Food-101. To provide a fair and unbiased comparison, all models were trained in the same experimental conditions, with the same preprocessing procedure, data augmentation methods, and training parameters.

The experimental findings suggest that EfficientNetB0 has the most stable and overall performance on the datasets and has an excellent potential to generalize the complicated visual patterns. ResNet50 also provided strong and consistent performance and has the advantage of deep residual architecture. In the meantime, MobileNetV2 reported competitive performance with high computational efficiency, which is why it is especially applicable to resource-constrained settings.

In general, the results indicate the significance of choosing suitable deep learning structures to enhance food image classification performance, particularly in scenarios in which accuracy and computational efficiency are factors of consideration in the real world.

List of Abbreviations

- | | |
|-------------------------------------|--|
| - Acc: Accuracy | - Food-11: Food-11 Image Dataset |
| - Adam: Adaptive Moment Estimation | - Food-34: Food Classification dataset |
| - AI: Artificial Intelligence | - Food-101: Food-101 Image Dataset |
| - ANN: Artificial Neural Network | - GPU: Graphics Processing Unit |
| - AUC: Area Under the Curve | - LR: Learning Rate |
| - CNN: Convolutional Neural Network | - ML: Machine Learning |
| - DL: Deep Learning | - mAP: Mean Average Precision |
| - F1-score: F1 Measure | - Prec: Precision |
| | - Rec: Recall |

- Res-VMamba: Residual Vision Mamba
- SGD: Stochastic Gradient Descent
- SHAP: SHapley Additive exPlanations
- SSM: State-Space Model
- TL: Transfer Learning
- ViT: Vision Transformer
- VMamba: Vision Mamba
- XAI: Explainable Artificial Intelligence

Introduction:

The study on food image recognition has become one of the important fields of computer vision and artificial intelligence because of the large number of applications in the fields of healthcare, nutrition management, agriculture, and smart food systems. The practical uses of auto-detecting food items by image include diet monitoring, calorie content determination, food quality and automated food recommendation. As the technology of deep learning evolves rapidly, image-based food classification has seen significant advances in the accuracy and strength of its use over the traditional computer vision-based methods [1].

Irrespective of these progresses, food image recognition is a difficult task. In most cases, there is high intra-class variance between food images due to variation in food ingredients, cooking techniques, presentation, lighting and camera angles. Simultaneously, similarity between different classes of food that are visually similar is another problem that makes the classification process even more complicated. Such

problems require the adoption of sophisticated deep learning implementations that are able to capture the fine-grained details and visual context of the object throughout the entire scene [2].

The recent research has had a growing trend in assessing and comparing the state-of-the-art deep learning models to ascertain their appropriateness in complex food image recognition tasks. Comparative analyses are used to find out the advantages and drawbacks of various architectures and support the choice of suitable models to be used in real-world contexts [3,8].

Research Problem:

The visual characteristics of food items are complex and therefore make the classification of food images difficult to do accurately. The food pictures are extremely different even within a similar category unlike the general object recognition tasks. The differences in portion size, cooking, texture, color and background have a significant impact on the visual appearance, which is not classified properly according to the traditional systems [1].

Moreover, different classes of foods have similar visual characteristics e.g., color, shape, and the texture that makes it more likely to be confused across classes. It is especially clear in fine-grained food recognition tasks, where subtle variations need to be found in order to differentiate between similarly looking dishes [2].

Even though the list of the proposed deep learning models that can be used to recognize food images is very long, existing papers

usually tend to concentrate on a few architectures or datasets. Lacking comparative performance between diverse architectures of the state-of-the-art, it is not easy to evaluate the scaling of models, their efficiency, and the ability to generalize [3].

Thus, the analysis of modern deep learning models and constructive comparison should be conducted to gain a better idea of their capabilities in food image classification tasks.

Importance of Research:

- Comparative evaluation framework:

The study provides a structured comparison of multiple state-of-the-art deep learning architectures under consistent evaluation criteria, enabling a fair and meaningful assessment of model performance.

- Performance analysis across architectures:

By examining different deep learning models, the research highlights how architectural design choices influence classification accuracy and robustness in food image recognition tasks.

- Architectural performance analysis:

The study investigates the architecture design options of various deep learning models and their effect on accuracy and resilience of classification handling food image recognition tasks.

- Lessons on scalability and efficiency:

It provides useful insights into the trade-off between the model size and computational efficiency and classification performance on top of the existing comparative studies in the

food and agricultural image classification markets .

- Practical relevance:

The findings can inform researchers and practitioners to use the right deep learning models in practical use, including nutrition evaluation, health assistance systems, and smart food administration systems.

Study Objectives:

- Compare the classification performance of the current deep learning models, such as convolutional neural networks and sophisticated architectures, to the food image datasets.
- Evaluate the capability of architectures in the visual complexity, inter-class similarity, and intra-class variation in food images.
- Compare models basing on their accuracy and on overall effectiveness, noting the trade-offs between model complexity and model performance.
- Offer recommendations on the appropriateness of various deep learning models to real-world food recognition problems, including dietary tracking, and smart food systems.

Theoretical Literature of the Study:

Food image recognition is one of the most vibrant research fields in the computer vision and artificial intelligence since it has extensive applications in healthcare, agriculture, nutrition surveillance, smart food systems, and intelligent store. Image-based

food recognition is essential in the dietary evaluation process, calorie computation, illness prevention, food quality management and elderly-friendly food diaries. As deep learning methods continue to rise at an alarming rate, the field of food image recognition has seen significant gains in performance and is now much more accurate, scalable, and robust than traditional image processing methods and machine learning models [2,13].

The first image recognition systems were based on manually designed feature extractions methods, in which handcrafted domain specialists designed the features manually according to their colour, texture and form. Although these techniques worked in limited situations, it could not cope with real world images because of change in lighting, background clutter, occlusion, and changes in viewpoint[9,12]. Food pictures are even more problematic due to the presence of high intra-class variation and an intense inter-class resemblance. Consequently, methods rooted in traditional methods could be hard to generalize over different types of food and datasets [4].

The advent of deep learning and in particular convolutional neural networks (CNNs) was a point in time in the history of image recognition research. Deep learning can automatically learn hierarchical representations with direct input of raw pixel data, which removes the requirement of manually engineered features[10]. This quality enables them to be specially applicable in food recognition, where appearances are multifaceted and highly adverse. It has been shown that deep

learning-based models are extremely effective in comparison with the traditional classification methods in various fields, such as food recognition and agricultural image analysis [5,11].

This chapter offers a summary of food image recognition models, especially deep learning models. It explains various types of models, such as generic convolutional neural networks, deep CNN models, and attention-enhanced CNN models, and upcoming state-space models. Moreover, the chapter also examines the commonly used food image datasets and presents their features and limitations, which becomes the basis of the experimental analysis in the subsequent chapters.

- **Models Used in Food Image Recognition**

The prevailing paradigm of food image recognition is that which is achieved through deep learning because of its strong capacity of gaining discriminative features of a complex visual information. Recent research has suggested and compared different deep learning architectures each with its unique pros and cons depending on the situation of application, the complexity of the data, and computational capabilities [6].

Applied to food image recognition, deep learning models may be divided into convolutional neural networks, deep CNN architectures, CNN architectures with attention mechanisms, and more recent ones, like state-space models. Comparative studies of these models have demonstrated that none of these architectures can be considered universal: the results significantly depend on

the size of the dataset, the number of classes, image quality, and the extent of visual similarity of food categories [2].

- **Convolutional Neural Networks (CNNs)**

Convolutional neural networks (CNNs) are a category of machine learning (ML) algorithm that uses a unique structure to classify images

Most food image recognition systems are based on convolutional neural networks. CNNs accept grid-like data (including images) and use convolutional filters which learn local spatial features. These filters allow CNNs to learn low-level features (edges, textures, and color gradients) in the early layers and more abstract and semantic features (object shape and structure) in the later layers [4].

Their hierarchical ability to learn features is a strength that CNNs have. Contrary to the traditional methods of feature design that involve a manual design of features, CNNs learn the features (relevant features) of data automatically through training. This renders them highly flexible to changes in food appearance due to differences in the method of cooking, ingredients as well as presentation styles. Consequently, CNN-based models are so common as a baseline to the food image classification task [2].

The use of CNNs in the classification of food images has been shown to be effective in a number of studies. As an illustration Suddul & Seguin (2023) compared various CNN models to classify foods and demonstrated that CNNs with the transfer learning and data augmentation strategies had high accuracy and F1-scores. Their findings emphasized the

significance of using the pretrained CNN models to enhance performance particularly in situations where the training data is small[4].

Even though the standard CNNs are successful, they have limitations in fine-grained classification problems. Classes of food frequently have a visual appearance that is alike, and hence it may be hard to differentiate between the classes using shallow or moderately deeply CNNs. This has been the impetus to create more intricate and sophisticated CNN architectures.

- **Deep CNN Structures**

Deep CNNs are based on the fundamental architecture of CNNs, which add depth to the network and introduce architectural advances to enhance learning efficiency and representation capability. Widely used deep CNN models have the following architectures: ResNet, DenseNet, EfficientNet, VGG, and Inception-based networks. The models are meant to resolve the difficulties of vanishing gradients, inefficient feature reuse, and high computational cost [5].

In the ResNet architectures, residual connections enable the information to traverse layers directly, making it possible to train very deep networks without its performance decreasing. DenseNet also enhances the reuse of features by linking each layer to all the other layers on feed-forward basis. The concept of EfficientNet proposes a compound scaling method which optimizes network depth, width, and resolution and obtains high accuracy at a relatively low computational cost [2].

Deep CNN models have always performed better than shallow models in food image recognition. A large-scale comparative study of 67 models of 16 deep learning architectures by Rasyidi et al. (2024) concluded that models based on EfficientNet attained the best accuracy on an Indonesian food dataset. They highlighted the trade-off between the model size and model performance where they indicated that deeper and complex models tend to give better accuracy, but at the expense of increased computation resource[2].

Equally, the efficacy of deep CNNs has been proven in studies in agricultural image analysis. Terzioğlu et al. (2025) also indicated that EfficientNet and ResNet-based models performed better than traditional machine learning methods in tomato disease detection tasks, which implies the scalability and strength of deep CNN models when used in a task that requires complex visual data[5].

- **CNN Models with Attention**

CNN architectures have been given attention mechanisms to help them to concentrate on the most informative areas of an image. Attention mechanisms in food image recognition can be used to make the models focus on pertinent visual features including food texture, food ingredients, and shape and reduce the background noise and other irrelevant information. This matters especially when there are genuine food photos, the backgrounds of which can be disorderly and incongruent [4].

Attention-enhanced CNN models transform the original CNNs by adding spatial, channelwise or multi-level attention

modules. These modules dynamically give weights to various regions or feature maps, which is the ability of the network to focus on important areas that add the most to the decisions made on classification. Due to this, attention-based models tend to be more accurate in fine-grained classification tasks than normal CNNs [2].

Comparative experiments have demonstrated that attention mechanisms enhance model robustness and particularly when it comes to differentiating two closely related food categories. According to Rasyidi et al. (2024), attention-enhanced architectures showed better performance with complex food datasets, but also reported that computational complexity increased[2]. This points out one important trade off between accuracy and efficiency in using attention-based models.

The attention mechanisms have also been utilized successfully in other image recognition fields. As an example, agricultural disease detection has been improved with attention-enhanced deep learning models that have exhibited better feature discrimination and classification in realistic settings [5]. The results indicate that attention CNNs can be effective in the image recognition of foods where high accuracy and discrimination are needed.

- **Models of State Space**

State-space models are a comparatively recent type of deep learning architecture which tries to capture long-range dependencies as well as global contextual information in a more efficient way than traditional CNNs. In comparison with CNNs,

which consider local spatial features mostly, state-space models are created to represent sequential and global interactions in data, and thus are applicable to intricate visual recognition problems [6].

Even though the use of state-space models has become popular in other fields like time-series analysis and natural language processing, use in image recognition remains relatively new. More recent studies propose that a combination of state-space modeling frameworks and deep learning frameworks can offer a wise trade off between local feature extraction and global context modeling. This applies especially to food image recognition, in which finer detail and general aesthetic structure matter to properly classify a food image [2].

Although the state-space models have so far not been exploited fully in food image recognition, they have potential benefits due to their ability to be widely scaled and lower costs to compute than very deep CNNs. State-space models have been found to be more significant in the next generation food recognition systems as more research is being done in this field.

Food Image Datasets

The food image datasets are a key aspect of food image recognition studies, as it directly affects the training of models, their evaluation, and ability to generalize. These properties of a dataset such as the number of classes, the quality of images, the accuracy of annotations, and the degree of visual diversity play a major role in model performance [4].

The initial food image datasets were small and less varied than is typical and therefore unable to support the creation of effective recognition systems. Newer activities have been directed towards developing bigger and more heterogeneous datasets to better capture the conditions of the real world. Rasyidi et al. (2024) proposed a large-scale Indonesian food dataset of more than 24,000 images of 160 types of foods, which focus on the absence of diversity in the literature[2]. Their contribution emphasized the need to have good quality and variety of datasets to enhance model generalization.

The other critical problem with food image collection is the data imbalance, with certain types of food represented in higher numbers than others and some with comparatively low sample sizes. To resolve this problem, data augmentation methods like rotation, flipping, scaling and color transformation are widely used. Suddul & Seguin (2023) showed that data augmentation contributed greatly to better classification and reduced overfitting of food image recognition models[4].

In addition to food recognition, knowledge obtained by other image-based areas also stresses on the significance of dataset design. Terzioğlu et al. (2025) demonstrated that datasets that were acquired in actual production scenarios resulted in stronger and more trusty deep learning models in agricultural disease detection[5]. The results are immediately relevant to food image recognition where the variability in the real world should be reflected in the training data.

To conclude, the research of food image recognition has to rely on the availability of the large, diverse, and well-annotated food

image datasets to further progress. Characteristics of the dataset not only affect the accuracy of the model but also define if trained models can be applied to real-life applications like healthcare, nutrition monitoring and smart food systems.

Review of Previous Studies:

Due to the rapid development of visual data and the proliferation of mobile devices, automatic food detection has become a major research topic in computer vision and artificial intelligence. Food image recognition has attracted considerable attention due to its importance in applications such as nutritional assessment, health monitoring, calorie estimation, and advanced food-related systems.

This chapter reviews previous studies on food image identification, focusing on the evolution of food recognition methods over time.

It begins with traditional computer vision techniques, then moves on to artificial intelligence-based techniques and their capabilities. A table comparing the different models used in previous research is then presented, followed by a detailed explanation of each model.

Deep Learning Models in Image Recognition of Food.

The comparative study of state-of-the-art deep learning architectures in food image classification presented by Rasyidi et al. (2024) utilized the Indonesian food as an object of study, where the visual diversity and complexity of the food are significant. The authors provided a novel dataset of 24,427

images belonging to 160 types of food categories, overcoming the shortcomings of the previous ones that did not have cultural and visual diversity. 67 models that rely on 16 modern architectures were tested, such as EfficientNet, ConvNeXt, or Swin Transformer. The performance indicated that EfficientNetV2-L was the most accurate with 85.44, which is effective in balancing the complexity of models and performance. This paper discussed the need to use big and diverse datasets and benchmarking of systems in order to create food recognition systems that are reliable enough to be applied to actual scenarios [2].

Reviews in detail of Deep Learning in Food Image Recognition.

Liu et al. (2025) used a review of deep learning methods in food image recognition to provide a general overview of the development of the area. The paper classified food recognition algorithms into the older methods of feature-based food recognition using traditional handcrafted algorithms and the more recent methods of food recognition using deep learning algorithms. It conducted a methodical survey of large food image datasets and mentioned their strengths and weaknesses on the criteria of size, diversity, and quality of annotation. Also, the authors compared the performances of different deep learning architectures on different datasets and also in different applications. Applied use cases (calorie estimation and food safety) and open challenges (dataset imbalance and difficulty in the fine-grained classification) were also discussed in the review. The study offers a useful theoretical background and contextual information to the comparative

analysis on the image recognition of food items [1].

Agricultural Image Classification Deep Learning Architectures.

Farahnakian et al. (2023) explored the efficacy of the state-of-the-art deep learning architectures to classify rice grains, which is a significant process in agricultural automation. Some of the latest models such as ResNet, VGG, EfficientNet and MobileNet were tested on a large dataset comprising 75,000 images of five types of rice. The findings indicated that EfficientNet was the most accurate in its classification of 99.67, whereas MobileNet had a better computational efficiency. It was also found that deep learning models are much more successful than the old fashioned machine learning techniques in addressing complex visual patterns and large scale data. Even though the study is agricultural, the findings on scalability, efficiency, and model choice would be very useful in food image recognition studies [3].

Object Detection Algorithms Comparative Analysis.

A comparative study of such popular deep learning-based image detectors as SSD, Faster R-CNN, and YOLO was conducted by Srivastava et al. (2021) with the help of the Microsoft COCO dataset. The models were compared by the study in terms of accuracy, precision, recall, and F1-score. The experimental findings showed that YOLOv3 was faster and more efficient than the other algorithms hence applicable in real time usage. The findings of the research are useful, although it did not concentrate on the

classification of food but on object detection, and they are useful in relation to the choice of a model depending on the requirements related to its use. The detection-based methods prove especially efficient in the context of food recognition, where several food objects are being displayed in one picture [6].

Agricultural Vision-based Deep Learning-based Feature Extraction.

Terzioglu et al. (2025) studied how deep learning-based feature extraction can be used with traditional machine learning classifiers to detect tomato diseases. The authors compared 21 architectures of deep learning and selected the top architectures like EfficientNet-B0 and ResNet-50 to extract features. Their findings revealed that the accuracy of deep feature combined with classical classifiers was greatly enhanced with a maximum of 92.0 accuracy. This combined method proved to be strong in actual production situation. Although the research was based on plant disease identification, its approach can be used to understand how deep learning can be efficiently implemented in challenging visual classification problems, which can be transferred to food image recognition systems [5].

Food Classification by Deep Learning with Data Augmentation.

Suddul & Seguin (2023) studied the use of deep learning and transfer learning methods to categorize food images based on a dataset of more than 16, 000 food images divided into 11 categories of food. Data augmentation techniques like rotation, flipping, and

zooming were used in the study to ensure that imbalance in the dataset and overfitting were overcome. The results of the experiment showed that EfficientNetV2 has a validation accuracy of 94.5% and F1-score of 94.7%. The authors have highlighted the significance of data augmentation and regularization methods in enhancing model generalization. This paper provides useful tips on how to improve performance using small or unbalanced datasets of food images [4].

NutriNet: Dietary Assessment Deep CNNs in the Early stage.

In their study, Mezgec & Koroušić Seljak (2017) came up with NutriNet, a deep convolutional neural network that was capable of recognizing the image of food and beverages when used in dietary assessment. The model was trained with a large dataset with more than 225,000 images and 520 categories of food and drinks. The classification accuracy of NutriNet was found to be 86.72% and it showed good performance in real-life smartphone images. The paper has discussed the limitations of food image recognition in non-controlled settings and the fact that deep learning is

better than the previous methods. NutriNet is one of the first large-scale food recognition applications of deep CNNs, providing the foundation of the further achievement in the domain [7].

Summary of Related Work

On the whole, the studies discussed in the present paper prove that deep learning has become the supreme method of food image recognition and other aesthetic identification tasks. Conventional machine learning methods are always beaten by convolutional neural networks, and specifically more advanced models, like EfficientNet. The quality data dataset, data augmentation and transfer learning are important in enhancing performance. Further research on the scalability and the robustness of deep learning models can be established through comparative studies in various fields including agriculture and object detection. All these findings are informative of the methodological decisions of the current research study and explain why it is necessary to conduct systematic performance comparison of state-of-the-art deep learning architectures in food image recognition.

Summary table

Table 1 Previous Studies summary

No.	Study	Domain / Task	Dataset	Models Used	Key Findings	Relevance to This Study
1	Rasyidi et al[2].	Food image classification	24,427 images, 160 Indonesian food classes	EfficientNet, ConvNeXt, Swin Transformer (67 models)	EfficientNetV2-L achieved highest accuracy (85.44%); highlighted trade-off between model size and performance	Provides large-scale benchmark and motivates comparative evaluation of modern architectures
2	Liu et al. [1]	Review of food image recognition	Multiple public food datasets	CNNs, attention models, deep architectures	Deep learning significantly outperforms traditional methods; dataset quality is critical	Offers theoretical foundation and identifies research gaps
3	Farahnakian et al. [3].	Rice grain classification	75,000 rice images (5 classes)	ResNet, VGG, EfficientNet, MobileNet	EfficientNet achieved 99.67% accuracy; MobileNet fastest	Demonstrates scalability and efficiency of deep CNNs
4	Srivastava et al [6].	Image detection	MS COCO dataset	SSD, Faster R-CNN, YOLOv3	YOLOv3 outperformed others in speed and efficiency	Supports use of deep learning for real-time visual tasks
5	Terzioğlu et al. [5].	Tomato disease detection	6,414 real-world agricultural images	EfficientNet, ResNet, DenseNet + ML classifiers	Hybrid deep feature extraction achieved 92% accuracy	Shows robustness of deep models in real environments
6	Suddul & Seguin [4].	Food image classification	16,000+ food images (11 classes)	CNNs, EfficientNetV2 (transfer learning)	Achieved 94.5% accuracy; data augmentation improved results	Highlights importance of transfer learning and augmentation
7	Mezgec & Koroušić Seljak [7].	Dietary assessment	225,953 images, 520 food & drink items	Custom deep CNN (NutriNet)	Achieved 86.72% accuracy; effective on smartphone images	Early large-scale application of deep CNNs in food recognition

Methodology:

The methodology comprises of a number of steps such as datasets selection, data exploration, data pre-processing, model selection, training setting and performance measurements. The stages are aimed at providing an opportunity to provide a comparison between models under controlled and uniform conditions.

Experimental Environment and Hardware Setup

To ascertain reproducibility and have a clear picture of the computation needs of this study, the experimentation was performed in a personal computing device and cloud computing.

A laptop with a clock speed of about 2.4 GHz, Intel Core i5 processor, and 8 GB memory was used as a local machine in the present study. This arrangement was primarily utilized to pre-process data, conduct exploratory data analysis and pre-test.

Cloud-based environments were used, namely Google Colab and Kaggle, because of the computational requirements of deep learning model training. These systems can access acceleration by the use of a GPU, which is much more effective in terms of training. Specifically, Google Colab provides the NVIDIA Tesla T4, which has around 12 GB of VRAM, and Kaggle offers the capabilities of similar GPUs based on the availability.

The experiments were carried out with the help of Python programming language, as well as deep learning frameworks such as TensorFlow and Keras. These platforms

allowed training models efficiently, computing with greater speed, and working with large-scale datasets of images.

It is also necessary to mention that the effectiveness of deep learning models, training time, and computational efficiency depends strongly on hardware resources. The GPUs installed in systems have a great potential of minimizing the time spent in training, as opposed to CPU-based systems. Such details will enable researchers in the future to repeat the experiments in similar conditions or modify the approach to their, at their disposal, computational resources.

Dataset Selection

Three publicly available benchmark datasets were chosen to test the performance of deep learning models under different degrees of complexity in classification. These datasets were well selected so as to reflect various degrees of difficulty in food image recognition tasks, and that is, low to high difficulty. The difference between datasets makes it possible to study in detail the influence of model performance on the number of classes, visual similarity between classes, and the size of the dataset.

Besides being used as training and evaluation material, the datasets offer a variety of real-life examples of food images, such as differences in light conditions, backgrounds, presentation styles, and image quality. This variety is critical in terms of testing the soundness and the capacity of generalization of deep learning models.

1. Food-11 Dataset

Food-11 dataset is a popular benchmark dataset in food image classification studies. It consists of 16,643 images, all of which

belong to 11 broad food classes, which is a rather low-complexity dataset.

Dataset Characteristics:

Type of data: RGB food images

Total number of images: 16,643 images

Number of classes: 11 categories

Form of image: JPEG (.jpg)

Format of file giving: ClassID ImageID.jpg.

Dataset size: 1.19 GB.

Food Categories:

The following classes are contained in the dataset:

Bread

Dairy products

Desserts

Eggs

Fried food

Meat

Noodles/Pasta

Rice

Seafood

Soup

Vegetables/Fruits

Data Structure:

The data is classified in three subsets:

Training set

Validation set

Evaluation (test) set

Such an a priori division guarantees the standard assessment and equal comparison of models.

Information Source and Use:

The data has been initially gathered and offered by the Multimedia Signal Processing Group (MMSPG) at EPFL. The main aim of developing this dataset was to aid in research in the field of food recognition and dietary analysis systems, especially in using this dataset in calorie estimation and food logging automated systems.

Visual Characteristics:

The pictures in this data set display:

Midrange fluctuation in light conditions.

Different camera angles

Diverse settings (restaurants, plates, and tables)

Practical styles of food presentation.



Normativity in This Research:

Food-11 can be used as a control dataset to test how well a model performs in a less complex classification task because of its small number of classes and somewhat defined boundaries between them.

2. Food-34 Dataset (Food Image Classification Dataset)

Food-34 dataset is a medium complexity dataset of about 24,000 images of 34 food categories. It contains a wide range of Indian as well as Western food products, which is appropriate to test the models on more diverse and culturally heterogeneous data.

Dataset Characteristics:

Type of data: RGB food images

Total number of images: ~24,000 images

Number of classes: 34 classes.

Image sources: Obtained on Google images.

Dataset size: 1.83 GB.

Data Composition:

The types of food included in the dataset are:

Fast food (burger, fries, sandwich)

Desserts (ice cream, donut, cheesecake)

Classic cuisine (butter naan, masala dosa, chole bhature)

A different number of images (around 200 to 1500) is present in each of the classes, creating imbalance between classes.

Data Source and Purpose:

Independent contributors on Kaggle curated the dataset with the goal of:

Supplying a variety of food imagery to train food classification models.

Research on menu recognition and menu recommendation.

Allowing real-life uses like automation of restaurants.

Visual Characteristics:

Great diversity of food presentation.

Mixed backgrounds (restaurants, homes, outside)

Differences in image quality and resolution.

Food diversity based on culture.

**Challenges:**

Categorical imbalance of classes.

High intra-class variability

Moderate inter-class similarity

Significance in this Research:

Food-34 can be classified as moderate, which is the best to test the extent to which models can handle greater diversity and complexity when contrasted with simpler datasets.

3. Food-101 Dataset

One of the most frequently used and difficult datasets in food image recognition is food-101. It contains 101,000 images, which are evenly spread across 101 food categories, and is therefore a high-complexity dataset.

Dataset Characteristics:

Data type: RGB pictures.

Total amount of images: 101,000 images.

The number of classes: 101 categories.

Images per class:

750 training images

250 test images

Image size: Rescaled to a maximum size of 512 pixels.

Data Source:

The data was generated by:

Lukas Bossard

Matthieu Guillaumin

Luc Van Gool

at ETH Zurich, and presented in 2014 (ECCV conference).

Data Collection Process:

The photos were obtained through Foodspotting.com, a social food photography application.

Users posted images of real world food in restaurants.

Naming was done according to the names of dishes.

Annotation:

Test images were checked manually.

There is purposeful noise in training images (mislabels, artifacts)

Purpose of Dataset:

The dataset was programmed to:

Give a standard of fine grained image classifications.

Test model discrimination between visually similar food items.

Replicate noisy data conditions.

Visual Characteristics:

Great within-class variability (food sameness with different appearances)

Great inter-class similarity (different foods look similar)

Real world (lighting, occlusion, backgrounds).

**Challenges:**

Training data label noise.

Fine-grained classification difficulty

Bias in the dataset in favor of Western cuisine.

Significance to the Present Research:

Food-101 is a high-difficulty benchmark enabling to evaluate the model performance under the conditions of complex real-world problems.

Exploratory Data Analysis

An exploratory analysis of the datasets was performed prior to the training of the deep learning models so as to get a feel of how the data is structured and what is the nature of the data.

In the exploratory analysis a number of critical aspects were examined:

To begin with, the overall amount of images in each dataset was examined. This allows knowing whether the size of the dataset used is adequate to train deep learning models.

Second, the dataset size of the classes was checked to learn the complexity of the classification.

Third, the spread of images among various classes was examined in order to determine the possible data imbalance issues. The class imbalance may have adverse impacts on the model performance in case certain classes have a significantly higher number of images. Also, random samples of each of the data set were manually examined to see changes in lighting conditions, camera angles, complexity of background, and food presentation styles. These differences are of significance in the challenge of the classification task.



This exploratory analysis assists in giving a better insight into the datasets and assists in the interpretation of the experimental findings later chapters.

Data Preprocessing

A series of preprocessing steps were performed to feed the images into the deep learning models before making sure that all the datasets are of the same format and training conditions.

To begin with, the size of all images was reduced to a constant number (224×224 pixels) which is the default size of delivered input to most pre-trained convolutional neural networks.



Second, the pixel values were scaled to the range of 0 to 1. This normalization is useful in making the training process more stable and better convergence in the optimization process.



Third, a validation split was used to split each dataset into training and validation sets. The model parameters are trained using the training set and the performance of the model is assessed using the validation set.

It is important to ensure a fair comparison between models because preprocessing has to be consistent across all datasets.

Data Augmentation

In order to enhance the generalization capacity of the models and minimize

overfitting, a number of data augmentation methods were used throughout the training.

The concept of data augmentation is an artificial mechanism that enhances the variety of training data with the help of creating distorted analogs of existing images.

The augmentation methods applied in this research are:

- Image rotation



- Horizontal flipping



- Slight zoom in and zoom out



These changes mimic effects of natural variations in image capture like varying camera angles and orientations.

Nonetheless, violent changes like random cropping were not done so as not to alter the primary contents of the food pictures.

Data augmentation can assist the models to learn stronger visual characteristics and have a better capacity to generalize on unseen information.

Model Selection

In this paper, three deep learning architectures were chosen to compare them:

- **ResNet50**

ResNet50 is a deep convolutional neural network that provides residual connections such that information flowing through the layers does not decline. This architecture allows training of very deep networks, but does not require the vanishing gradient problem.

- **MobileNetV2**

MobileNetV2 is intended to be lightweight. It is especially applicable to computing systems that have scarce computation capability like mobile devices. In various image classification problems, it is also capable of competing with other products that are relatively lightweight.

- **EfficientNetB0**

EfficientNet proposes a scaling strategy of a network that balances between network depth, width and resolution of input. This technique will enable EfficientNet models to be very accurate with a comparatively low computational cost.

The choice of these three models is due to the fact that they are various design philosophies within **the field of deep learning architecture:**

- a profound residual network (ResNet)
- a mobile network architecture (MobileNet)
- a compressed balanced architecture (EfficientNet)

It is with this diversity that meaningful comparison of performance would be possible between various architectural strategies.

Training Configuration

The training configuration was the same in all models to make the comparison in the experiment fair.

The settings were the following:

- Optimizer: Adam
- Loss: Categorical Cross-Entropy.
- Evaluation Metric: Accuracy
- Batch Size: 32

The applicability of the same training parameters is important to make sure that variations in performance are more because of the architecture of the models and less because of the training conditions.

Evaluation Metrics

The results of each model were measured in relation to a number of standard classification measures.

The primary measure of evaluation considered in the current study is the classification accuracy as a ratio of the correctly classified images.

Also, it is possible to use additional measures like precision, recall, and F1-score, which would allow a more complete assessment of the model performance, particularly in the case of imbalanced class distributions.

The application of several criteria of evaluation can give a better evaluation of the strengths and weaknesses of each model.

Results:

Experimental Setup

The experiments were done on the basis of a deep learning model that was written in Python and used the libraries TensorFlow and Keras. The base models were pre-trained convolutional neural networks, which were

trained using the ImageNet dataset by transfer learning.

All the images were also scaled to 224 x 224 pixels to fit the size of the input required by the models used. Normalization of pixel values was done to enhance stability of training. Rotation and horizontal flipping data augmentation methods together with some minor zoom were used to enhance diversity in the data set and minimize overfitting.

A similar validation split was used to separate each dataset into a training and validation set. It was also ensured that the training setup was the same in all the models so that the comparison only based on the architectural differences but not on the differences in training parameters.

The main training parameters used in the experiments are shown in Table 2.

Table 2 Training Configuration

Parameter	Value
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Evaluation Metric	Accuracy
Batch Size	32
Image Size	224 × 224

It is necessary to maintain the same training conditions in order to compare the performance of models of different architectures fairly.

Results on the Food-11 Dataset

The initial experiments have been conducted in Food-11 dataset, which is a somewhat simple image classification task as there is a small variety of food categories.

Table 3 summarizes the classification accuracy achieved by the three deep learning models on the Food-11 dataset.

Table 3 Model Performance on Food-11 Dataset

Model	Training Accuracy	Validation Accuracy
ResNet50	0.9937	0.9239
MobileNetV2	0.9272	0.9093
EfficientNetB0	0.9912	0.9207

The results show that ResNet50 had the highest training accuracy of the compared models, then EfficientNetB0, and MobileNetV2 had slightly lower training accuracy.

Nevertheless, MobileNetV2 proved to be a good performer based on its lightweight structure as it exhibited competitive performance with a much lower computational cost. This renders it quite appropriate to be used in resource constrained systems like mobile and embedded systems.

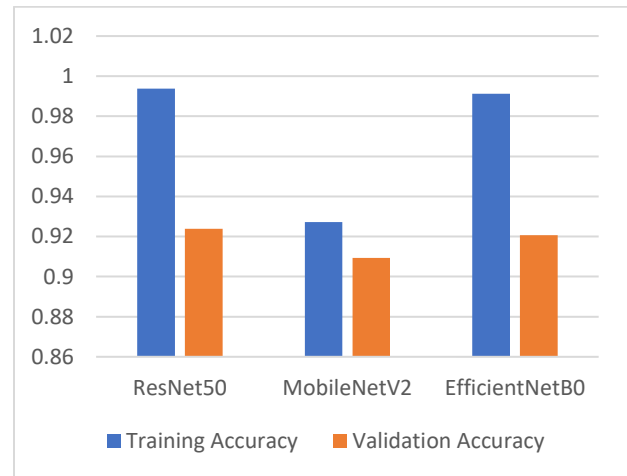
EfficientNetB0 was also highly trained, as it has the advantage of its strategy of scaling networks based on compound scaling that balances network depth, width, and resolution.

ResNet50 had the highest training accuracy, but this can be explained by the fact that it has a more elaborate architecture and a bigger

model capacity to support more complex feature representations. Nonetheless, this complexity can be difficult to balance, and can even be more costly in terms of computation than simpler architectures.

Figure 1 illustrates the comparison of model accuracy on the Food-11 dataset.

Figure 1 Comparison of Model Accuracy on Food-11



These findings indicate that more efficient architectures like EfficientNet can increase their accuracy even when using relatively small data sets because of balanced scaling of the network depth, width, and resolution.

Nevertheless, it is necessary to mention that the accuracy of validation is lower than that of the training accuracy in all models. This conduct reflects the existence of generalization issues, which are typical of visual similarity image classification exercises between food groups.

Results on the Food-34 Dataset

The second experiment was done on the basis of the Food-34 dataset which involves more difficulty in the classification because of the

number of categories of food and the visual similarity of classes.

Table 4 summarizes the results obtained from training the models on the Food-34 dataset.

Table 4 Model Performance on Food-34 Dataset

Model	Training Accuracy	Best Validation Accuracy
EfficientNetB0	0.9864	0.9557
ResNet50	0.9861	0.9525
MobileNetV2	0.9850	0.9525

The results of the experiment show that EfficientNetB0 has the best validation accuracy (95.57%), which proves its high capacity to generalize to unseen data. ResNet50 and MobileNetV2 were also competitive with both achieving a highest validation accuracy of 95.25%.

Even though the training accuracy of all models was very high, EfficientNetB0 performed marginally better on validation performance, indicating the usefulness of its compound scaling approach to the representation of complex visual patterns.

MobileNetV2 showed impressive results even in the lightweight format, with an accuracy level equal to that of deeper models and less computational power. This is what makes it especially applicable to implementation in resource-constrained environments.

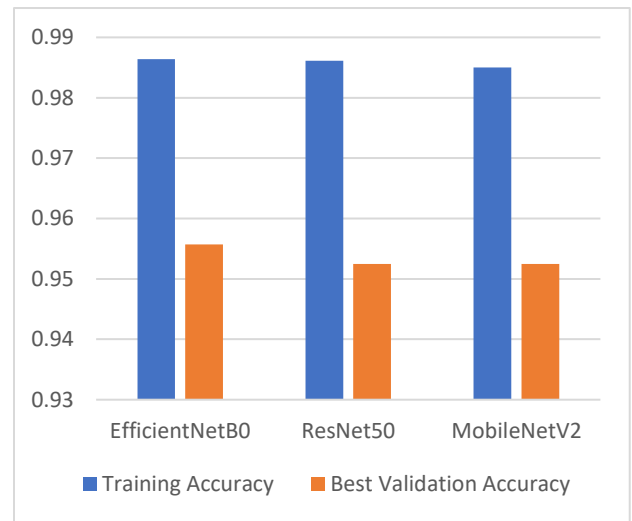
Strong results were also achieved by ResNet50 because it has a deep residual architecture that allows efficient learning of

complex features representations. Nonetheless, its increased level of computation can restrict its efficiency with more streamlined architecture like EfficientNet and MobileNet.

On the whole, the findings support the claim that the current deep learning architectures are capable of high accuracy in food image classification tasks, with EfficientNetB0 being the most efficient in terms of accuracy and efficiency.

Figure 2 shows the comparison of training accuracy across models for the Food-34 dataset.

Figure 2 Model Accuracy on Food-34 Dataset



The findings affirm that the accuracy of the classification is usually reduced with the intricacy of the dataset.

Results on the Food-101 Dataset

The last series of tests was done with Food-101 data that is one of the most difficult tasks in food image recognition.

Food-101 dataset has many classes of food where the visual similarity between them is

very high. A lot of food products have a similar texture, color, and shape so it is especially hard to classify them using fine-grained classification.

Table 5 summarizes the observed model performance.

Table 5 Model Performance on Food-101 Dataset

Model	Training Accuracy	Best Validation Accuracy
EfficientNetB0	0.9864	0.9494
ResNet50	0.9899	0.9367
MobileNetV2	0.9144	0.8892

The outcome of the experiment proves that there are distinguishable performance differences between the tested models. The highest validation accuracy (94.94) was obtained by EfficientNetB0, which means that it has better generalization capacity than other models. This is due to its compound scaling strategy, which is able to balance depth, width, and resolution of the network.

ResNet50 was also able to perform well, recording a validation accuracy of 93.67%. It is highly capable of learning complex visual features based on its deep residual architecture; though it marginally fail to compete with EfficientNetB0 on generalization.

On the contrary, MobileNetV2 was less successful in performance with a validation accuracy of 88.92. Nonetheless, it still is a computationally efficient model, as it is lightweight that makes it applicable to be

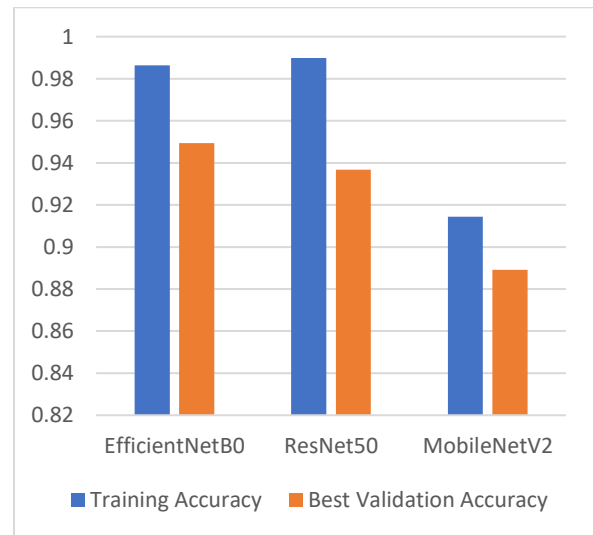
deployed in resource-constrained environments.

Moreover, it is possible to notice that both EfficientNetB0 and ResNet50 had very high training accuracy and lower validation one, which suggests that there was a mild overfitting. This should be the case in deep learning models that are trained on comparatively complicated data.

In general, the findings support the claim that EfficientNetB0 is the best trade-off, whereas MobileNetV2 has a trade-off of being efficient at the expense of a marginally lower level of performance.

Figure 3 shows the approximate comparison of model performance on the Food-101 dataset.

Figure 3 Model Accuracy on Food-101 Dataset



The findings indicate that fine-grained food classification tasks are challenging and that it is essential to choose suitable deep learning architectures.

Comparative Analysis of Model Performance

To better understand the behavior of the models across datasets, Table 6 presents a consolidated comparison of training accuracy across all experiments.

Table 6 Overall Model Comparison

Model	Food-11 (Val Acc)	Food-34 (Val Acc)	Food-101 (Val Acc)
ResNet50	0.9239	0.9525	0.9367
MobileNetV2	0.9093	0.9525	0.8892
EfficientNetB0	0.9207	0.9557	0.9494

Discussion of Experimental Findings

This part provides a detailed comparative study of the performance of the reviewed deep learning models on datasets of different complexity, i.e., Food-11, Food-34, and Food-101. The comparison is mainly based on the accuracy of validation because this is the ability of the models to generalize to unseen data.

The findings indicate that all the models performed well in all the datasets but evident differences are observed with increase in dataset complexity. All models performed similarly in the Food-11 dataset, which is a relatively low-complexity classification task with ResNet50 performing slightly better than the others. This implies that even traditional deep architectures can be effective where there is low class diversity.

The validation accuracy of EfficientNetB0 (95.57%), ResNet50 (95.25%), and MobileNetV2 (95.25%) are close to each other as the complexity of the Food-34 dataset increases. It implies that, in terms of more diverse and complex visual features, the compound scaling strategy of EfficientNet offers a slight advantage.

In the most difficult case, the Food-101 data, where the classification is fine-grained with a high degree of inter-class similarity, the EfficientNetB0 model obviously outperforms other models, with the highest validation accuracy (94.94%). ResNet50 was found to be performing well but slightly worse than EfficientNetB0, whereas MobileNetV2 was considerably less accurate (88.92%).

These results demonstrate the influence of architecture on model performance. EfficientNetB0 had always been found to have an excellent generalization ability on all the datasets especially in complicated scenes as a result of its balanced scaling of depth, width, and resolution. ResNet50 demonstrated a high and consistent performance, which is advantageous due to its deep residual learning structure, though with a higher computational complexity level.

MobileNetV2, on the other hand, despite being slightly lower in its accuracy on complex data, showed competitive performance based on their lightweight nature. This renders it especially useful in real world applications where computational resources are constrained, or in mobile and embedded systems.

Moreover, it is found that all the models showed greater training accuracy compared to validation accuracy, which means that there was slight overfitting. This is a common characteristic of deep learning models, particularly in tasks that involve complex and fine-grained image classification.

In general, the comparative study confirms that EfficientNetB0 offers the best trade-off between accuracy and generalization performance, whereas MobileNetV2 follows a beneficial balance in terms of efficiency and performance, and ResNet50 offers good performance with more computation demands.

Summary of Experimental Results

It was found that with ease of use (on simpler datasets like Food-11) all models were highly accurate. The difference in performance was however more pronounced with an increase in the complexity of datasets.

The best overall performance was obtained with EfficientNetB0 on complex datasets, and the next best performance was obtained with ResNet50 with good and consistent performance. MobileNetV2 was not as accurate but efficient because of its lightweight architecture.

Also, a low distance between training and validation accuracy demonstrates slight overfitting. EfficientNetB0 is overall more accurate and generalized, although MobileNetV2 fits better in resource-limited scenarios.

Conclusion and Future Work

Conclusion

Food image recognition has become a significant field of research in computer vision because of its broad use in the healthcare field, diet monitoring, nutritional management and intelligent food systems. As the methods of deep learning have advanced, the results of automated image classification systems have greatly enhanced since the use of the convolutional neural networks.

The main aim of this work was to compare three popular deep learning architectures in food image recognition ResNet50, MobileNetV2, and EfficientNetB0. These models are a contrasting design philosophy of deep residual, lightweight mobile, and scaling of compound models.

In order to compare models fairly, each model was trained under the same conditions, with the same preprocessing methods, data augmentation methods, and training parameters. The assessment was performed on three datasets of the growing levels of complexity: Food-11, Food-34 and Food-101.

The experiment showed that EfficientNetB0 performed the best overall, especially with more challenging datasets and, therefore, has a high generalization and the capacity to render deeper visual patterns. ResNet50 also provided a stable and high performance, whereas MobileNetV2 could also provide competitive results at a much lower computational cost, and it is well in line with resource-constrained settings.

Moreover, it was observed that the performance of the classification declines

with the complexity of the dataset. This can be largely explained by the fact that the number of classes is greater, and the visual similarity of the various classes of foods is high, which complicates the task of classification.

On the whole, this paper establishes that deep learning models are extremely useful in food image recognition tasks today. The comparative analysis reveals useful information to choose the suitable models depending on the nature of the trade-off between the accuracy and the computational efficiency based on the application needs.

Future Work

- Comparing the performance of other deep learning architecture models like Vision Transformers (ViT) or ConvNeXt and Swin Transformer to those of CNN-based architectures.
- Improving model performance using hyperparameter optimization methods: by increasing the learning rate, learning batch size, optimizer parameters, and training schedules.
- Enhancing the classification accuracy and robustness of various models by ensemble learning, which involves using multiple models to combine predictions and refines the model.
- Adding attention mechanisms so that the models can concentrate on the most informative areas of the food images and enhance performance of fine-grained recognition.
- Using explainable artificial intelligence (XAI) methods, e.g.,

Grad-CAM or SHAP, to gain more insight into deep learning models to make classification decisions.

- enlarging datasets with a greater number of food categories and life pictures having different light conditions and backgrounds.
- Designing real-time food recognition systems that can be implemented in mobile applications to monitor their diets, estimate calories and assist them in personalized nutrition.
- Investigating the concept of multi-task learning that integrates food classification with similar tasks (e.g. detecting the ingredients, estimating portions size, and calculating the calories).

References list

- [1] Liu, D., Zuo, E., Wang, D., He, L., Dong, L., & Lu, X. (2025). Deep learning in food image recognition: A comprehensive review. *Applied Sciences*, 15(14), 7626. <https://doi.org/10.3390/app15147626>
- [2] Rasyidi, M. A., Mardhiyyah, Y. S., Nasution, Z., & Wijaya, C. H. (2024). Performance comparison of state-of-the-art deep learning model architectures in Indonesian food image classification. *Bulletin of Electrical Engineering and Informatics*, 13(5), 3355–3368. DOI: <https://doi.org/10.11591/eei.v13i5.7996>
- [3] Farahnakian, F., Sheikh, J., Farahnakian, F., & Heikkonen, J.

- (2023).A comparative study of state-of-the-art deep learning architectures for rice grain classification. *Journal of Agriculture and Food Research*, 13, 100890. <https://doi.org/10.1016/j.jafr.2023.100890>
- [4] Suddul, G., & Seguin, J. F. L. (2023).A comparative study of deep learning methods for food classification with images. *Food and Humanity*, 1, Article 100018. <https://doi.org/10.1016/j.foohum.2023.07.018>
- [5] Terzioğlu, H., Gölcük, A., Shakarji, A. M. A., & Al-Bayati, M. Y. (2025).Comparative analysis of deep learning-based feature extraction and traditional classification approaches for tomato disease detection. *Agronomy*, 15(7), 1509. <https://doi.org/10.3390/agronomy15071509>
- [6] Srivastava, S., Divekar, A. V., Anilkumar, C., Naik, I., Kulkarni, V., & Pattabiraman, V. (2021).Comparative analysis of deep learning image detection algorithms. *Journal of Big Data*, 8, Article 66. <https://doi.org/10.1007/s12536-021-00066-6>
- [7] Mezgec, S., & Koroušić Seljak, B. (2017).NutriNet: A deep learning food and drink image recognition system for dietary assessment. *Nutrients*, 9(7), 657. <https://doi.org/10.3390/nu9070657>
- [8] Chikatimarla, A. K. (2024). A comparative analysis of state-of-the-art deep learning architectures for image recognition tasks. *Integral Research*, 1(9). <https://doi.org/10.1016/j.intres.2024.100009>
- [9] Ciocca, G., Micali, G., & Napoletano, P. (2020). State recognition of food images using deep features. *IEEE Access*, 8, 32003–32017. <https://ieeexplore.ieee.org/abstract/document/8998272/>
- [10] Ma, P., Li, C., Rahaman, M. M., Yao, Y., Zhang, J., Zou, S., Zhao, X., & Grzegorzec, M. (2023). A state-of-the-art survey of object detection techniques in microorganism image analysis: From classical methods to deep learning approaches. *Artificial Intelligence Review*, 56, 1627–1698. <https://doi.org/10.1007/s12536-023-00000-0>
- [11] Nuthakki, S., Buddiga, S. K. P., & Koganti, S. (2024). Exploring deep learning models for image recognition: A comparative review. *Signal & Image Processing: An International Journal*, 15(3). <https://doi.org/10.5121/sipij.2024.15301>
- [12] Rasheed, F., Dar, A. A., Ahmad, S., & Hussain, G. (2026). Comparative analysis of deep state-of-the-art CNN architectures over Food-101 dataset

using transfer learning and the impact of fine tuning. In *Advancements in Generative Artificial Intelligence* (pp. 243–259). Springer. [Comparative Analysis of Deep State-of-the-Art CNN Architectures Over Food-101 Dataset Using Transfer Learning and the Impact of Fine Tuning | Springer Nature Link](#)

- [13] Salim, N. O. M., & Mohammed, A. K. (2024). Comparative analysis of classical machine learning and deep learning methods for fruit image recognition and classification. *Traitement du Signal*, 41(3). <https://doi.org/10.18280/ts.410322>