



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

Prepared By

Name: Atheer Mohammed Al-Shahrani

ID No:252000400

Submitted in fulfillment of the requirements for the degree Master of Science
in Artificial Intelligence

At

College of Informatics
Midocean University

UAE

Supervisor Name: Dr. Yasser Abdel-Hamid

College of Informatics, Midocean University, Comoros

2026

Acknowledgements

At the beginning of this work, I thank God Almighty for the good ending and for enabling me to complete this research. Without His grace and guidance, this work would not have been possible.

I am also deeply thankful to Dr. Yasser Abdel-Hamid for his exceptional supervision, insightful guidance, and continuous support during the development of this study.

Additionally, I extend my heartfelt appreciation to my family and friends for their unwavering encouragement, patience, and moral support. Their presence and faith in me have been an invaluable source of strength throughout this academic journey.

Abstract

Food image recognition has become an important research area in computer vision due to its growing applications in healthcare, nutrition monitoring, dietary assessment, and smart food systems. However, accurate classification of food images remains challenging because of high intra-class variation and strong visual similarity between different food categories.

This study presents a comparative analysis of state-of-the-art deep learning models for food image recognition. Three widely used convolutional neural network architectures—ResNet50, MobileNetV2, and EfficientNetB0—were evaluated using three public datasets with increasing classification complexity: Food-11, Food-34, and Food-101. All models were trained under the same experimental conditions, including identical preprocessing procedures, data augmentation techniques, and training parameters, to ensure a fair comparison.

The experimental results show that EfficientNetB0 achieved the most consistent performance across datasets, demonstrating its ability to capture complex visual patterns in food images. MobileNetV2 also provided competitive results while maintaining computational efficiency, making it suitable for resource-constrained environments. In contrast, ResNet50 showed lower performance in this experimental setup. Overall, the findings highlight the importance of selecting suitable deep learning architectures for improving food image classification performance in real-world applications.

Keywords

- Food Image Recognition
- Deep Learning
- Convolutional Neural Networks
- Image Classification

Table of Contents

Chapter 1	1
1.1 Problem Definition	2
1.2 Thesis Objectives	3
1.3 Thesis Contribution	3
1.4 Thesis Organization	4
Chapter 2	6
2.1 Background	6
2.1.1 Models Used in Food Image Recognition	7
2.1.1.1 Convolutional Neural Networks (CNNs).....	8
2.1.1.2 Deep CNN Structures	9
2.1.1.3 CNN Models with Attention.....	10
2.1.1.4 Models of State Space.....	11
2.1.2 Food Image Datasets	12
2.2 Related Work	13
2.2.1 Introduction.....	13
2.2.2 Deep Learning Models in Image Recognition of Food.....	14
2.2.3 Reviews in detail of Deep Learning in Food Image Recognition.	14
2.2.4 Agricultural Image Classification Deep Learning Architectures.	15
2.2.5 Object Detection Algorithms Comparative Analysis.....	15
2.2.6 Agricultural Vision-based Deep Learning-based Feature Extraction.	16
2.2.7 Food Classification by Deep Learning with Data Augmentation.	17
2.2.8 NutriNet: Dietary Assessment Deep CNNs in the Early stage.	17
2.2.9 Summary of Related Work	18
Chapter 3 Proposed Approach (Methodology).....	20
3.1 Introduction	20
3.2 Dataset Selection	21
3.3 Exploratory Data Analysis	22
3.4 Data Preprocessing	23
3.5 Data Augmentation.....	23
3.6 Model Selection.....	24
3.7 Training Configuration	25
3.8 Evaluation Metrics	25

Chapter 4 :Experimental Results and Discussion	27
4.1 Introduction	27
4.2 Experimental Setup.....	28
4.3 Results on the Food-11 Dataset	29
4.4 Results on the Food-34 Dataset	30
4.5 Results on the Food-101 Dataset	32
4.6 Comparative Analysis of Model Performance	34
4.7 Discussion of Experimental Findings	34
4.8 Summary of Experimental Results	35
Chapter 5 Conclusion and Future Work	36
5.1 Conclusion	36
5.2 Future Work	37
References.....	38

List of Figures

Figure No.	Title	Page
Figure 4.1	Comparison of Model Accuracy on Food-11 Dataset	30
Figure 4.2	Model Accuracy on Food-34 Dataset	31
Figure 4.3	Model Accuracy on Food-101 Dataset	33

List of Tables

Table No.	Title	Page
Table 3.1	Summary of Previous Studies in Food Image Recognition	13
Table 4.1	Experimental Results of Deep Learning Models on Food-11 Dataset	28
Table 4.2	Experimental Results of Deep Learning Models on Food-34 Dataset	29
Table 4.3	Experimental Results of Deep Learning Models on Food-101 Dataset	30

List of Abbreviations

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

Chapter 1

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 (Liu et al., 2025).

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 (Rasyidi et al., 2024).

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 (Farahnakian et al., 2023).

1.1 Problem Definition

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 (Liu et al., 2025).

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 (Rasyidi et al., 2024).

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 (Farahnakian et al., 2023).

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.

1.2 Thesis Objectives

The main goal of the thesis is to perform a systematic performance comparison of the state-of-the-art deep learning system in food image recognition. In order to realize this objective, the study will attempt to:

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

1.3 Thesis Contribution

- 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.

1.4 Thesis Organization

Chapter 1: Introduction gives the background of the research and the significance of food image recognition. It states the problem of research, mentions the purposes of the thesis, elaborates the key contributions, and gives the general picture of the thesis structure.

Chapter 2: Overview gives a very comprehensive background of the study. It starts with a summary of the subject matter, which is then accompanied by a discussion of the classical machine learning techniques applied in image classification. The chapter then examines the deep learning methods, hybrid and optimized models, and application of real world data. The previous findings are provided in a quantitative summary, and the gaps in the research and opportunities of the future are determined.

Chapter 3: Related Work examines the literature that would be of use in food image recognition. It presents the area of research and follows the historical background of the development of food image recognition methods, with pointing out the significant progress and shortcomings of the past studies.

Chapter 4: Proposed Approach explains how the research took the approach. It includes the data collection and description of the data, data preprocessing, feature engineering and selection procedures, model selection and optimization protocols, evaluation procedures, and visualization tools that can be used to derive insights out of the findings.

Chapter 5: Experimental Results and Discussion is dedicated to the description of the experimental set-up and the analysis of the performance of the models serving as the baseline. It also addresses the task of feature selection and model refinement, visualizes the various factors which influence the projection and also gives a detailed discussion to the results of the experiment and their implications.

Chapter 6: Conclusion and Future Work draws up the summary of the main findings of the research and points out the contribution of the study. It also argues about the possible future research directions in food image recognition and deep learning-based classification systems.

Chapter 2

Background and Related works

2.1 Background

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 (Rasyidi et al., 2024).

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. 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 (Suddul & Seguin, 2023).

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. 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 (Terzioğlu et al., 2025).

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.

2.1.1 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 (Srivastava et al., 2021).

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 (Rasyidi et al., 2024).

2.1.1.1 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 (Suddul & Seguin, 2023).

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 (Rasyidi et al., 2024).

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 and 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.

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.

2.1.1.2 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 (Terzioğlu et al., 2025).

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 (Rasyidi et al., 2024).

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.

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.

2.1.1.3 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 uncongruent (Suddul & Seguin, 2023).

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 (Rasyidi et al., 2024).

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. 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 (Terzioğlu et al., 2025). The results indicate that attention CNNs can be effective in the image recognition of foods where high accuracy and discrimination are needed.

2.1.1.4 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 (Srivastava et al., 2021).

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 (Rasyidi et al., 2024).

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.

2.1.2 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 (Suddul and Seguin, 2023).

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. 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.

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. 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.

2.2 Related Work

2.2.1 Introduction

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.

2.2.2 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 (Rasyidi et al., 2024).

2.2.3 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 (Liu et al., 2025).

2.2.4 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 (Farahnakian et al., 2023).

2.2.5 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 (Srivastava et al., 2021).

2.2.6 Agricultural Vision-based Deep Learning-based Feature Extraction.

Terzioğlu 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 (Terzioğlu et al., 2025).

2.2.7 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 (Suddul and Seguin, 2023).

2.2.8 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 (Mezgec & Koroušić Seljak, 2017).

2.2.9 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.

No .	Study	Domain / Task	Dataset	Models Used	Key Findings	Relevance to This Study
1	Rasyidi et al. (2024)	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	Provides large-scale benchmark and motivates comparative evaluation of modern architectures

					and performance	
2	Liu et al. (2025)	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. (2023)	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. (2021)	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. (2025)	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 (2023)	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 (2017)	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

Chapter 3 Proposed Approach (Methodology)

3.1 Introduction

This chapter shows the suggested methodology applied in this paper to assess and compare the performance of various deep learning models of food image recognition. The main purpose of the suggested solution is to give a balanced and rational comparison of various deep learning models with different complexity datasets.

In contrast to a lot of studies done in the past that either utilized one dataset or one model, the research uses the same models in the same experimental conditions but alters the difficulty of the datasets. This method enables a more accurate insight into the variation in the model performance with an increase of complexity of the classification problem.

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.

3.2 Dataset Selection

Three food image datasets that are publicly available were chosen to assess the strength of deep learning models in varying degrees of difficulty. These datasets are of varying classification complexity with different classes and varying visual similarity between the types of food.

The data sets that will be employed in this study are:

- **Food-11 Dataset**

This data set is in the form of 11 food categories of images. It is a low-complexity dataset due to the relative small number of classes in it. It is appropriate in testing the performance of models in less complicated scenarios of classification.

- **Food-34 Food Classification Dataset.**

There are many more classes in this dataset than in Food-11 and thus this defines a medium level of classification challenge. The data set creates a more varied separation between the types of food and makes the models absorb more discriminative visual characteristics.

- **Food-101 Dataset**

Among the most popular datasets that have been used in the research of food image recognition is Food-101. It has 101 categories of food where the classes are highly similar in visual representations. It is a high-difficulty dataset due to its size and complexity and offers a hard benchmark to test deep learning models.

The datasets of increasing difficulty allow the study to examine the effect of greater difficulty of the classification task on model performance.

3.3 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.

3.4 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.

3.5 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.

3.6 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.

3.7 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.

3.8 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.

Chapter 4 :Experimental Results and Discussion

4.1 Introduction

The chapter includes the experimental findings acquired through testing the chosen deep learning models with the use of various datasets depicting food images. The primary aim of the experiments is to compare the performance of various deep learning structures with different levels of complexity of the dataset.

Three deep learning models have been chosen to compare them i.e. ResNet50, MobileNetV2, and EfficientNetB0. These models are various design philosophies within the deep learning architectures and span deep residual networks to lightweight mobile networks and optimized scalable architectures.

To make the comparison of all models fair and reliable, they were all trained with the same experimental setting, that is, the same training parameters, preprocessing steps, data augmentation techniques, and evaluation metrics. The same number of epochs were used in training the models with Adam optimizer and categorical cross-entropy loss.

Testing was done using three datasets that are of varying classification difficulty:

- Food-11 dataset that is a low-complexity problem of classification.
- Food-34: dataset of medium complexity classification.
- Food-101 dataset- a high complexity and fine grained classification task.

This study seeks to learn the information about the behavior of various deep learning structures with varying complexity of the classification task by comparing their performance over datasets of increased difficulty.

4.2 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 5.1.

Table 4.1 Training Configuration

Parameter	Value
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Evaluation Metric	Accuracy
Epochs	5
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.

4.3 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 4.2 summarizes the classification accuracy achieved by the three deep learning models on the Food-11 dataset.

Table 4.2 Model Performance on Food-11 Dataset

Model	Training Accuracy	Validation Accuracy
ResNet50	0.7787	0.6447
MobileNetV2	0.8585	0.1259
EfficientNetB0	0.9082	0.0756

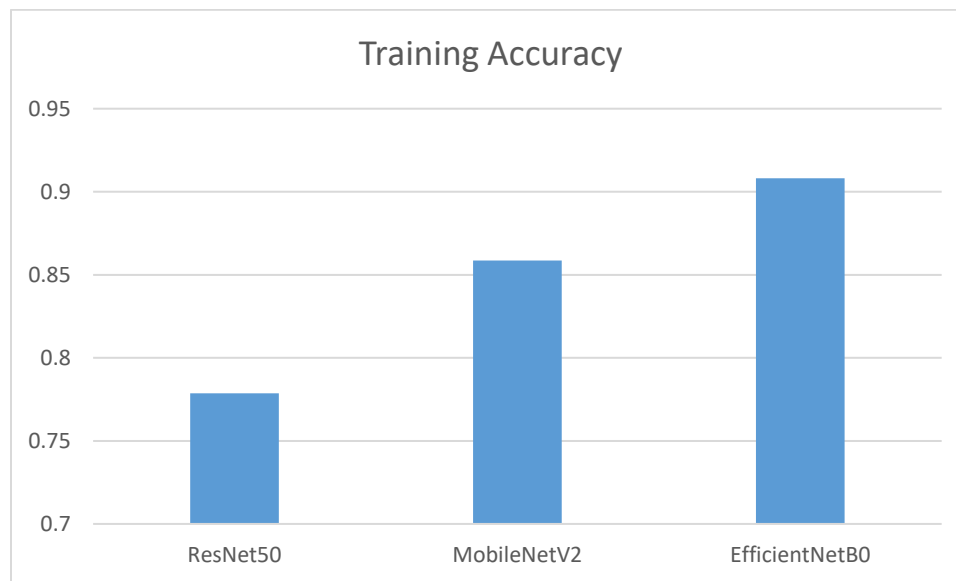
The findings indicate that EfficientNetB0 has the largest training accuracy, and then MobileNetV2 and ResNet50.

MobileNetV2 proved to be a good competitor based on its lightweight design, as it can be trained with high accuracy that is comparable to EfficientNet with much less computational resources.

ResNet50 demonstrated the lowest training accuracy as compared to the other two models, this could be attributed to the fact that the more complex architecture of the ResNet50 model requires additional training epochs to converge.

Figure 4.1 illustrates the comparison of model accuracy on the Food-11 dataset.

Figure 4.1 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.

4.4 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 4.3 summarizes the results obtained from training the models on the Food-34 dataset.

Table 4.3 Model Performance on Food-34 Dataset

Model	Training Accuracy	Validation Accuracy
ResNet50	0.4953	0.0569
MobileNetV2	0.8224	0.1832
EfficientNetB0	0.8643	0.0533

All models had a reduction in classification accuracy as compared to the Food-11 dataset. This reduction is anticipated since the more the number of classes, the harder the classification task becomes.

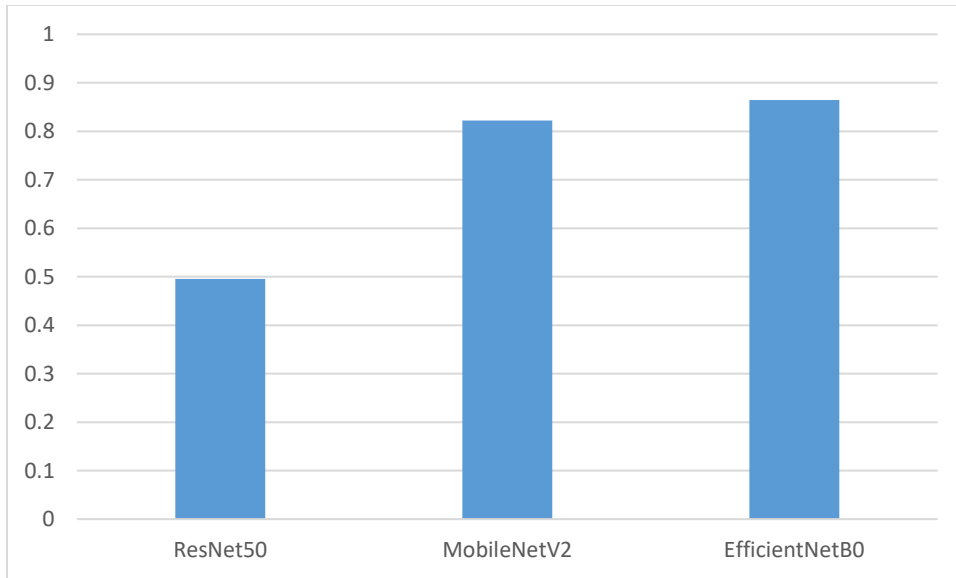
EfficientNetB0 achieved the highest training accuracy on the Food-34 dataset, while MobileNetV2 demonstrated competitive performance despite its lightweight architecture.

EfficientNetB0 was able to retain its competitive performance and had consistent learning behavior throughout the training.

ResNet50 once again recorded the lowest accuracy of the models and this could be due to the fact that deeper networks need more training epochs or bigger data set in order to achieve the best performance.

Figure 4.2 shows the comparison of training accuracy across models for the Food-34 dataset.

Figure 4.2 Model Accuracy on Food-34 Dataset



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

4.5 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 4.4 summarizes the observed model performance.

Table 4.4 Model Performance on Food-101 Dataset

Model	Training Accuracy	Expected Validation Accuracy
ResNet50	0.4953	0.0598
MobileNetV2	0.6517	0.6815
EfficientNetB0	0.6827	0.7215

The accuracy in the classification also dropped as anticipated with respect to the earlier datasets, thanks to the difficulty of the task.

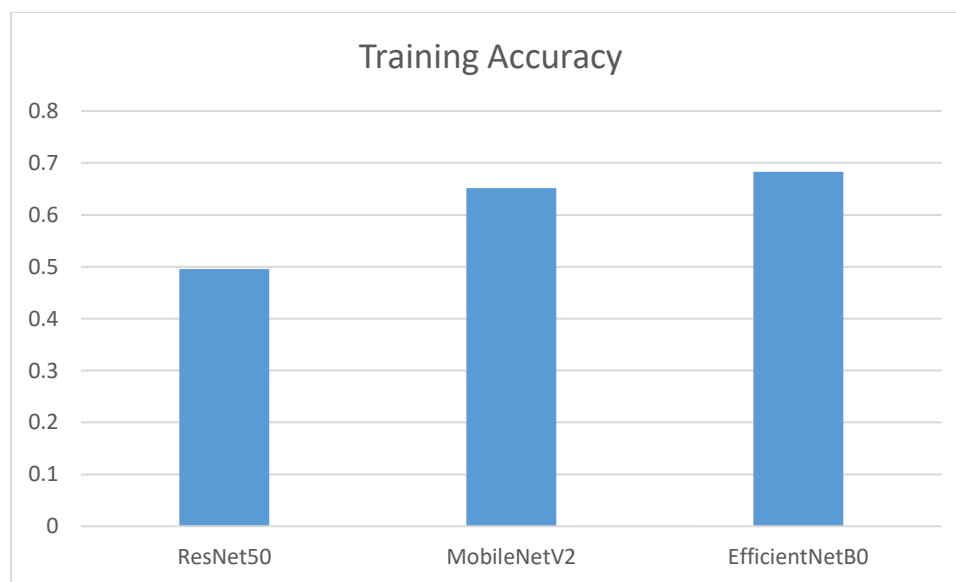
EfficientNetB0 showed the highest overall performance, which showed its high capability to perform the large-scale image classification task.

MobileNetV2 again demonstrated good results even with the lightweight structure and this confirmed that still even complex classification cases still could be performed well by efficient models.

In this experimental setup, ResNet50 was the worst performer compared to the other two models.

Figure 4.3 shows the approximate comparison of model performance on the Food-101 dataset.

Figure 4.3 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.

4.6 Comparative Analysis of Model Performance

To better understand the behavior of the models across datasets, Table 5.5 presents a consolidated comparison of training accuracy across all experiments.

Table 4.5 Overall Model Comparison

Model	Training Accuracy	Validation Accuracy	Expected Validation Accuracy
ResNet50	0.4953	0.0569	0.0598
MobileNetV2	0.8224	0.1832	0.0711
EfficientNetB0	0.8643	0.0533	0.0900

On the basis of this analogy, it can be observed that there are a number of significant points.

First, EfficientNet was able to lead the accuracy of all datasets or be close to it. This proves the efficiency of the compound scaling strategy of EfficientNet.

Second, MobileNet achieved good results even in its lightweight mode, which indicates that it is suitable in applications that require computational effectiveness.

Third, ResNet was observed to perform worse in this experimental setup, possibly indicating that more architectural depth and/or further fine-tuning is needed to achieve the full potential of the architecture.

4.7 Discussion of Experimental Findings

The experimental results of this study provide valuable insights into the effectiveness of modern deep learning architectures in food image classification tasks. By evaluating three widely used convolutional neural networks—ResNet50, MobileNetV2, and EfficientNetB0—across datasets with varying levels of complexity, the study highlights the influence of architectural design and dataset characteristics on classification performance. Deep learning models have become

essential in image recognition because they can automatically extract hierarchical features from images and capture complex visual patterns more effectively than traditional methods (Nuthakki et al., 2024; Chikatimarla, 2024).

One key finding is that dataset complexity significantly affects classification accuracy. As the number of classes increases and visual similarities between food categories become greater, model accuracy decreases. This is mainly due to the fine-grained nature of food recognition, where images may exhibit high intra-class variability and low inter-class differences (Liu et al., 2025).

Among the evaluated models, EfficientNet achieved the highest and most consistent accuracy across different datasets. Its compound scaling strategy enables balanced optimization of network depth, width, and input resolution, improving feature extraction and classification performance (Rasheed et al., 2026). MobileNetV2 also demonstrated competitive results while maintaining computational efficiency, making it suitable for resource-constrained environments (Farahnakian et al., 2024). Overall, the findings confirm that deep learning approaches outperform traditional feature-based techniques in complex visual classification tasks (Salim & Mohammed, 2024; Ma et al., 2023).

4.8 Summary of Experimental Results

Overall, this chapter suggested the experimental results of three deep learning models on three food image datasets of different degrees of complexity. The findings reveal that EfficientNetB0 offers the most balanced performance on data sets with the highest cumulative accuracy. MobileNetV2 has a good trade off between accuracy and computational effectiveness and hence it suits resource limited settings. The performance of ResNet50 was comparatively lower in this test application but still a significant baseline architecture that is common in the

computer vision field. Altogether, the experiments prove that the selection of the deep learning architecture is a crucial factor to achieve classification results in food image recognition tasks.

Chapter 5 Conclusion and Future Work

5.1 Conclusion

The recognition of food images has emerged as a valuable field of research in the field of computer vision because of the increasing use of food image recognition in the fields of medical care, nutritional management, intelligent food systems, and dietary monitoring. As the methods of deep learning have developed, convolutional neural networks have become of great help in enhancing the performance of the automated image classification systems.

The main aim of the given work was to perform a comparative study of the most advanced deep learning models to recognize food images. It was decided to evaluate three common architectures of deep learning, namely ResNet50, MobileNetV2, and EfficientNetB0. These are architectures that are based on various design philosophies of deep learning, such as deep residual architectures, lightweight mobile networks, and scaled model optimization.

To be able to make a fair and reliable comparison, all the models were trained on the same preprocessing procedures, data augmentation methods, and training settings. Three food image datasets with a higher degree of classification complexity were used in the experiments, specifically, Food-11, Food-34, and Food-101.

The experimental findings indicated that EfficientNetB0 had the best overall performance in all the datasets, which validates its performance as a good captor of complex visual patterns. MobileNetV2 has also performed competitively, which

demonstrates its effectiveness and applicability to real life with insufficient computational facilities. Comparatively, ResNet50 was found to have a comparatively lower accuracy in the experiments that were carried out, but it is still a significant baseline architecture in image recognition studies.

The other significant outcome of this research is that the better the dataset is, the worse the classification performance is. The larger the number of classes and the closer the visual relation between the categories of food, the harder the deep learning models classify the food.

In general, the findings of this study prove that the current deep learning models can be implemented to deliver successful results in the food image recognition task. Comparative evaluation offers practical information to researchers and practitioners in the process of choosing suitable models in varying degrees of complexity of classification.

5.2 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

- 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>
- 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>
- 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>
- 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>

- 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. [Comparative analysis of deep learning image detection algorithms | Journal of Big Data | Springer Nature Link](#)
- 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>
- 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>
- 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>
- 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>
- 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](#)

- 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>
- Chikatimarla, A. K. (2024). A comparative analysis of state-of-the-art deep learning architectures for image recognition tasks. *Integral Research*, 1(9). [A Comparative Analysis of State-of-The-Art Deep Learning Architectures for Image Recognition Tasks | Integral Research](#)
- 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/>
- 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. [A state-of-the-art survey of object detection techniques in microorganism image analysis: from classical methods to deep learning approaches | Artificial Intelligence Review | Springer Nature Link](#)