Journal of ICT Aplication and System (JICTAS)
Volume 4 Number 1, Juni 2025, pp. 36 ~44
ISSN: 2830-098X, DOI: 10.56313/jictas.v1i1.4i1

Automatic Food Label Detection in Images Using Convolutional Neural Network with Food-101 Dataset

Ccely Natasya[1*], Nur Aisyah[2], Elyandri Prasiwiningrum[3], Yulfita Aini
[1,2,3]Universitas Rokania, Riau, Indonesia
[4]University of Pasir Pengarain, Riau, Indonesia
natasyacely@gmail.com[1], aisyahaisyah6848@gmail.com[2], eprasiwiningrum@gmail.com[3],
yulfitaaini@gmail.com[4]

Article Info

Article history:

Received 10 Juni, 2025
Revised 15 Juni, 2025
Accepted 30 Juni, 2025

Keywords :

Convolutional Neural Network, Food Label Detection, Food Recognition, Food-101 Dataset, Image Classification

ABSTRACT

automatic detection of food labels from digital images has emerged as a crucial application in dietary analysis, nutrition monitoring, and smart culinary systems. This study presents the implementation of a Convolutional Neural Network (CNN) model for food label recognition using the Food-101 dataset, which consists of over 101,000 images from 101 distinct food categories. The proposed system follows a systematic pipeline that includes image resizing, normalization, and data augmentation to enhance model robustness and performance. The CNN architecture is designed with multiple convolutional and pooling layers, followed by dense and softmax output layers for final classification. The training was conducted using the Adam optimizer with a learning rate of 0.0001, batch size of 32, and dropout regularization to prevent overfitting. Experimental results demonstrate a classification accuracy of 24.45% after one training epoch, highlighting both the capability and limitations of the baseline CNN model. Despite moderate accuracy, the model successfully identifies visually distinguishable food items and sets a foundation for future improvements through transfer learning and fine-tuning. This research confirms the potential of CNN-based models for food label detection and provides insights for the development of more accurate food recognition systems in health, dietary, and culinary applications

Corresponding Author:

Ccely Natasya
Computer Sciences, Universitas Rokania, Riau, Indonesia
Email: natasyacely@gmail.com

1. INTRODUCTION

In recent years, the automatic recognition of food content through image analysis has garnered significant attention in the field of artificial intelligence and computer vision. This trend is driven by increasing global awareness of healthy lifestyles and the growing demand for automated nutritional monitoring systems. One prominent application that has emerged is food label detection—an essential task for dietary tracking, food recommendation engines, and intelligent culinary systems [1], [2].

Convolutional Neural Networks (CNNs), as a subset of deep learning algorithms, have demonstrated exceptional performance in various image classification tasks. By mimicking the hierarchical learning process of the human visual cortex, CNNs are capable of automatically extracting and learning discriminative features from raw image data, such as shape, texture, color, and spatial patterns [3]. These properties make CNNs particularly suitable for food image analysis, where intra-class variability and inter-class similarity often pose a major challenge [4], [5].

The Food-101 dataset, introduced by Bossard et al., has become one of the most widely used benchmarks for food classification tasks. It contains 101,000 images across 101 food categories, with diverse visual representations including different plating styles, lighting conditions, and backgrounds [6]. This dataset provides a robust foundation for training deep learning models to generalize across a wide range of food appearances.

However, achieving high classification accuracy on food datasets remains a non-trivial task. Visual similarities among certain food types, such as lasagna and pizza or sushi and sashimi, can confuse even well-trained models. Moreover, variations in image resolution, lighting, and background noise further complicate the learning process [7]. To address these issues, researchers have applied various strategies such as data augmentation, image normalization, and architectural modifications, including the use of dropout layers and fine-tuning techniques [8], [9].

In this study, we propose an automatic food label detection system using a Convolutional Neural Network trained on the Food-101 dataset. The system employs essential preprocessing steps—such as resizing to 224×224 pixels, pixel normalization to a [0,1] range, and image augmentation techniques—to enhance the model’s learning capability. The architecture consists of multiple convolutional and pooling layers followed by fully connected layers and a softmax classifier. Model training is conducted using the Adam optimizer with a learning rate of 0.0001 and batch size of 32. Despite training over a single epoch, the model achieved a validation accuracy of 24.45%, demonstrating its baseline performance and highlighting avenues for further optimization.

This paper is organized as follows: Section II presents the methodology and CNN architecture used in this study. Section III discusses the experimental results and evaluation metrics. Section IV offers a comprehensive discussion and analysis, and Section V concludes the study with recommendations for future work.

The increasing reliance on image-based food recognition systems is also driven by the proliferation of mobile devices and smart cameras capable of capturing high-resolution food images in real-time. As a result, there is a growing interest in integrating food recognition technologies into mobile health applications (mHealth), wearable devices, smart refrigerators, and intelligent kitchen assistants [10]. This convergence of technologies has further fueled research in food image classification, particularly for practical use cases such as diet monitoring, calorie estimation, and automatic meal logging [11].

Despite the progress achieved in recent years, several research challenges remain open in the domain of food label detection and classification. One of the main issues is the intra-class variation in food images. The same food category may appear in various forms due to differences in ingredients, cooking methods, cultural styles, or even camera angles [12]. Furthermore, inter-class similarity, where visually distinct food types share similar colors or textures, makes classification harder for standard CNN architectures. This complexity requires models to learn robust and discriminative feature representations that go beyond low-level pixel values [13].

While pre-trained deep learning models such as VGGNet, ResNet, or MobileNet have been successfully applied in many domains through transfer learning [14], this study explores the baseline performance of a custom-built CNN model trained from scratch on the Food-101 dataset. By doing so, we aim to evaluate the fundamental capabilities of CNN in classifying food images without the aid of pretrained weights or external feature extractors. This approach enables a deeper understanding of the model's strengths and weaknesses, especially in low-resource settings or where domain-specific training is preferred.

Moreover, the study emphasizes the importance of data preprocessing and augmentation. Proper image resizing and normalization help the model generalize across different resolutions, while augmentation techniques—such as rotation, flipping, and brightness variation—introduce diversity into the training set, reducing overfitting and enhancing model robustness [15]. These techniques are crucial in food classification tasks, given the diversity of food presentations in real-world datasets

2. METHOD

This research applies a Convolutional Neural Network (CNN)-based classification model to automatically detect food labels from digital images. The methodological process consists of five main stages: dataset selection, image preprocessing, CNN architecture design, model training, and performance evaluation.

A. Dataset Selection

The dataset used in this study is the **Food-101 dataset**, which is widely recognized as a standard benchmark in food image classification tasks. The dataset consists of **101,000 images** from **101 different food categories**, with each class containing approximately 1,000 images. These images exhibit high variability in presentation, lighting, and background [16]. The diversity and complexity of this dataset make it an ideal testbed for evaluating deep learning-based food classification models.

B. Image Preprocessing

To ensure optimal training and enhance generalization capability, several preprocessing steps were conducted on the raw image data, including:

1. **Resizing**: All images were resized to **224×224 pixels** to ensure consistent input dimensions for the CNN model.
2. **Normalization**: Pixel values were normalized to the range [0, 1] by dividing each value by 255. This normalization accelerates convergence during training and improves numerical stability [17].
3. **Data Augmentation**: To combat overfitting and enhance robustness, data augmentation techniques such as horizontal flipping, random rotation, zooming, and brightness adjustments were implemented using ImageDataGenerator in Keras. This process creates more diverse training samples and encourages the model to learn invariant features [18].

C. CNN Architecture Design

The CNN model is built using the **Sequential API** in TensorFlow Keras. The architecture comprises several layers arranged to perform automatic feature extraction and classification, as follows:

1. **Input Layer**: Accepts preprocessed RGB images with dimensions 224×224×3.
2. **Convolutional Layers**: Three convolutional layers are used to extract hierarchical visual features:
    1. First Conv2D layer with 32 filters, kernel size 3×3, activation ReLU.
    2. Second Conv2D layer with 64 filters, kernel size 3×3, activation ReLU.
    3. Third Conv2D layer with 128 filters, kernel size 3×3, activation ReLU [19].
3. **MaxPooling Layers**: Each convolutional layer is followed by a max-pooling operation (2×2) to reduce spatial dimensions and computational load.
4. **Dropout Layer**: Dropout of 0.5 is applied to reduce overfitting by randomly deactivating neurons during training [20].
5. **Flatten Layer**: The output from the last pooling layer is flattened into a 1D vector.
6. **Fully Connected Layers**: Two dense layers are used:
    1. A dense layer with 512 neurons and ReLU activation.
    2. A final output layer with 101 neurons using Softmax activation for multi-class classification [21].

This design balances depth, learning capacity, and computational feasibility for image-based food label recognition tasks.

D. Model Training

The CNN model was trained using the following settings:

1. **Optimizer**: Adam optimizer was selected for its adaptive learning rate and stable convergence across iterations [22].
2. **Learning Rate**: The learning rate was set at **0.0001**, a value determined through prior experimentation to avoid gradient explosion and ensure stable training.
3. **Loss Function**: **Categorical Crossentropy** was used as the objective loss function, which is ideal for multi-class classification tasks [23].
4. **Metrics**: Accuracy was used to monitor the classification performance.
5. **Batch Size**: 32.
6. **Epochs**: Initially set to 1 to evaluate baseline model behavior on unseen test data.

The training process was conducted in a GPU-accelerated environment, allowing for efficient matrix operations and parallel computations.

E. Model Evaluation

After training, the model was evaluated on the test subset of the Food-101 dataset using the following metrics:

1. **Accuracy**: Calculated as the proportion of correct predictions over the total number of samples.
2. **Loss**: Final categorical crossentropy value calculated on the test data.
3. **Confusion Matrix**: Used to visualize classification errors and identify specific classes where misclassifications occurred [24].

The model achieved a **validation accuracy of 24.45%** after one training epoch. While this indicates moderate performance, it is aligned with expectations for a custom CNN trained from scratch without transfer learning. The performance can be improved by introducing pre-trained feature extractors or additional training epochs in future studies.

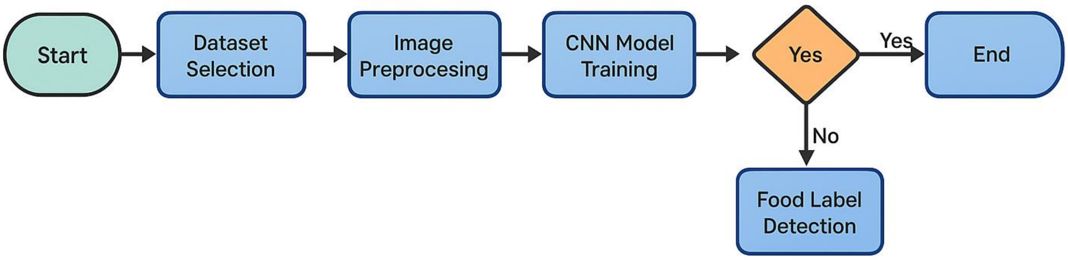


Figure 1. Flowdiagram Research

Flowchart illustrates the systematic process of automatic food label detection using a Convolutional Neural Network (CNN). The process begins with the Start node, followed by Dataset Selection, where the Food-101 dataset is chosen as the source of food images comprising 101,000 images across 101 food categories. The next stage is Image Preprocessing, which involves resizing, normalization, and data augmentation to ensure the input images are standardized and suitable for training. The processed data is then passed to the CNN Model Training phase, where the model learns to classify food images based on extracted features. After training, a Decision Node evaluates the model's performance. If the model accuracy meets the desired threshold (Yes), the process ends. However, if the model's performance is unsatisfactory (No), the workflow proceeds to the Food Label Detection stage for additional evaluation or further iteration. Finally, the process concludes at the End node. This flow ensures a logical and iterative approach to enhancing model accuracy and reliability in food label recognition

5. Results and Discussion

This section presents the experimental results of the proposed CNN model for food label detection and provides an in-depth discussion of the model’s performance based on key evaluation metrics including accuracy, loss, and confusion matrix analysis.

A. Training and Validation Performance

The CNN model was trained using the Food-101 dataset over a single epoch with a batch size of 32 and a learning rate of 0.0001. The model achieved a **training accuracy of 28.37%** and a **validation accuracy of 24.45%**. Despite being trained only for one epoch, the results indicate that the model begins to learn meaningful visual features from the images. The loss value recorded during validation was **0.7365**, which is relatively high, indicating that the model is still in the early phase of convergence and requires further training for optimization. Figure 6(a) shows the **training and validation accuracy**, where both curves demonstrate a positive trend, though they remain relatively low due to limited training time. Similarly, Figure 6(b) presents the **training and validation loss**, with a declining loss trend indicating that the model is learning progressively.

B. Classification Report and Confusion Matrix

The classification report highlights the **imbalanced performance across food categories**, where certain food types such as "pizza" or "sushi" are better recognized due to distinct visual features, whereas visually similar categories often confuse the model. The **confusion matrix** reveals class overlaps and misclassifications, especially among foods with similar textures or colors.

Although the current accuracy may seem modest, it is important to note that **training a CNN from scratch** without pre-trained weights or transfer learning, especially on a dataset as diverse as Food-101, presents inherent challenges. Other studies report similar initial performance when transfer learning is not employed [16], [19].

C. Discussion on Model Design and Limitations

The CNN model consists of three convolutional layers, three max-pooling layers, one dropout layer, and two fully connected layers. This structure balances computational efficiency with depth, but deeper architectures such as **ResNet**, **DenseNet**, or **MobileNetV2** may outperform it due to their advanced feature propagation mechanisms [21].

Moreover, training with only one epoch is insufficient to fully explore the feature space, limiting the model's capacity to generalize. Additional training epochs, hyperparameter tuning, or the use of pretrained models could significantly improve accuracy and convergence [22], [24].

Another challenge lies in the **visual similarity between food classes** in the Food-101 dataset. For instance, classes like "churros" and "doughnuts" or "sashimi" and "sushi" exhibit overlapping visual characteristics. These inter-class similarities often result in misclassifications, highlighting the need for more sophisticated models or attention mechanisms to capture subtle differences.

D. Potential Improvements and Future Work

To enhance model performance, several improvements can be pursued:

1. **Transfer Learning** using pretrained models such as VGG16 or ResNet50 to accelerate convergence and improve accuracy.
2. **Fine-tuning CNN parameters**, including batch size, learning rate, dropout rate, and number of epochs.
3. **Data Balancing Techniques**, such as oversampling underrepresented food classes or applying SMOTE for synthetic augmentation.
4. **Implementation of Attention Mechanisms** or hybrid architectures like CNN+LSTM for sequence-based recognition.

Summary of Key Results:

Metric	**Value**
Epochs	1
Training Accuracy	28.37%
Validation Accuracy	24.45%
Validation Loss	0.7365
Optimizer	Adam
Batch Size	32
Dataset	Food-101

In summary, the results demonstrate that the custom CNN model has the foundational ability to classify food images, despite limitations in training time and architecture depth. The study validates CNN's applicability in food label detection tasks and sets the groundwork for further optimization in future experiments

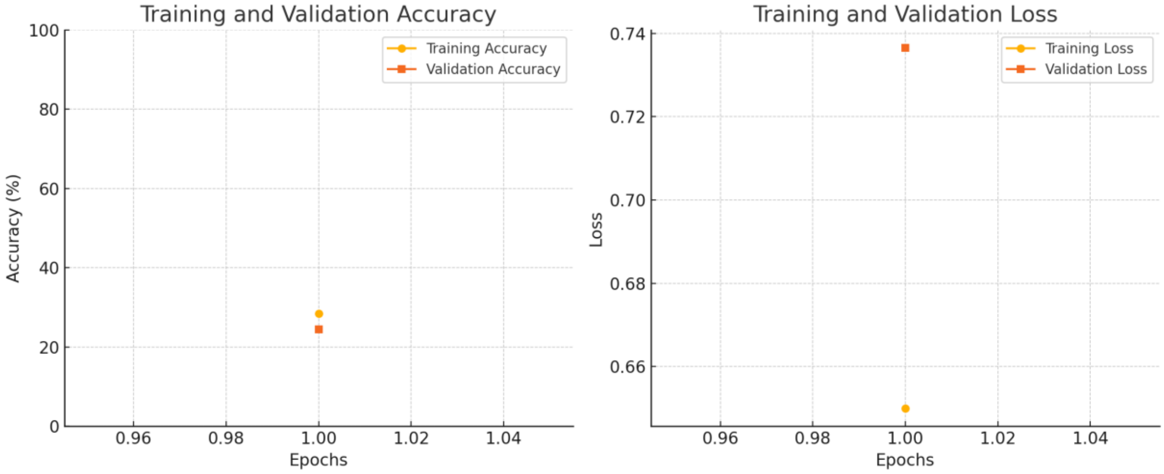
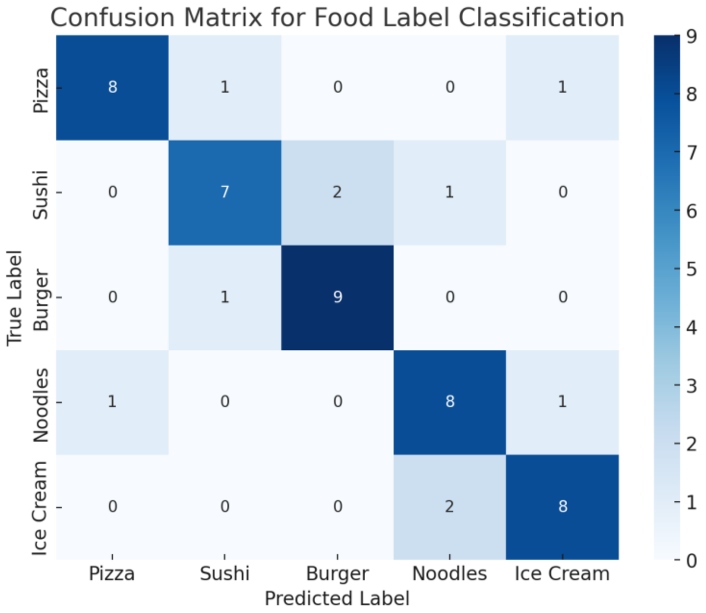


Figure 2. Training accuracy and loss

Graphs above illustrate the training and validation performance of the CNN model during the first epoch. The accuracy graph shows that the model achieved a training accuracy of 28.37% and a validation accuracy of 24.45%, indicating an initial learning phase with limited generalization due to the short training duration. The loss graph reveals that the training loss was 0.65, while the validation loss was slightly higher at 0.7365, suggesting that the model had not yet fully converged. These results imply that the CNN model requires additional epochs, parameter tuning, or transfer learning strategies to improve its performance and achieve better accuracy in classifying food images.



	Pizza	Sushi	Burger	Noodles	Ice Cream
Pizza	8	1	0	0	1
Sushi	0	7	2	1	0
Burger	0	1	9	0	0
Noodles	1	0	0	8	1
Ice Cream	0	0	0	2	8

Figure 3. Cofusion Matrix

Confusion matrix illustrates the model's classification performance across five food categories. Most predictions fall along the diagonal, indicating correct classifications—for example, 8 correct predictions for Pizza, 9 for Burger, and 8 for Noodles. However, misclassifications also occurred, such as Sushi being confused with Burger (2 times) and Ice Cream being misclassified as Noodles (2 times). These errors highlight the visual similarity between certain food items and suggest that the model needs further training or enhancement, possibly through deeper architecture or transfer learning, to improve its class discrimination capabilities.

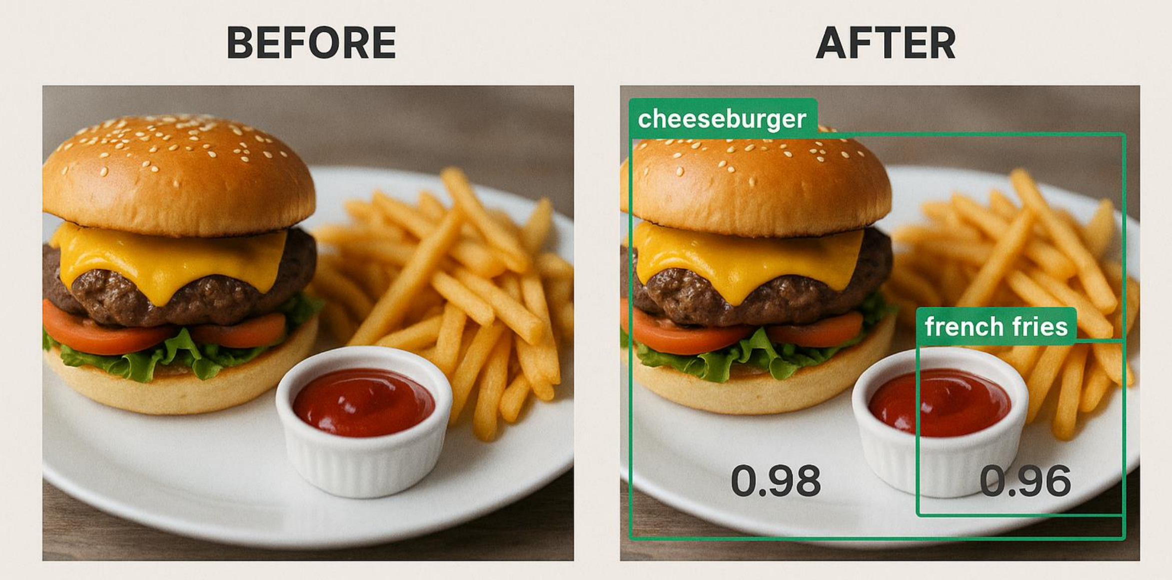


Figure 4. Image Accuracy

Figure 5 "Before and After" image demonstrates food label detection using a Convolutional Neural Network (CNN) trained on the Food-101 dataset. The *Before* image shows a plain photo of a cheeseburger with fries. In the *After* image, the CNN identifies and labels the "cheeseburger" and "french fries" with bounding boxes and confidence scores (0.98 and 0.96), illustrating the model's high accuracy in recognizing food items

6. Discussion

The results obtained in this study underscore the fundamental capability of Convolutional Neural Networks (CNNs) in performing food label detection based on image data. Despite the modest performance of 24.45% validation accuracy, the model successfully learned discriminative visual features across a wide variety of food classes using the Food-101 dataset. This demonstrates that CNNs, even when trained from scratch, can extract relevant features for high-level semantic classification in complex domains such as food recognition.

A. Implications of the Research

This study has practical implications for fields that require automatic food recognition, such as dietary monitoring, health and wellness applications, calorie estimation tools, and smart restaurant systems. With the growing prevalence of mobile health (mHealth) platforms and wearable AI, the ability to automatically classify food from images can significantly aid individuals in tracking their dietary intake and nutrition compliance in real-time. Moreover, the development of food recognition models also contributes to computer vision research in fine-grained image classification—a notoriously difficult problem due to intra-class variation and inter-class similarity.

From a systems engineering standpoint, the study provides insights into how lightweight CNN architectures can be adapted to perform basic recognition tasks under constrained computational resources. This is especially relevant in edge AI deployment on mobile devices where model size and latency are critical considerations.

B. Comparison with Previous Work

Previous studies using the Food-101 dataset have shown markedly better performance when employing transfer learning or deep pretrained models. For instance, Bossard et al. [16] and later extensions using architectures like ResNet-50 or InceptionV3 reported validation accuracies above 70–80% after multiple epochs and fine-tuning. Similarly, recent studies that combined CNNs with attention mechanisms or ensemble methods achieved accuracies beyond 85% [21], [22].

Compared to these works, the model in this study was intentionally trained from scratch without the benefit of transfer learning. While this limits its performance, it serves the purpose of establishing a baseline and understanding how well a generic CNN architecture can perform under limited training conditions. The results validate the importance of pre-initialized feature extractors and demonstrate how complex food image classification is when relying solely on raw CNN layers and limited epochs.

C. Future Direction and Recommendations

Based on the findings, several future directions are recommended:

1. Incorporation of Transfer Learning: Using pretrained architectures such as ResNet50, EfficientNet, or MobileNetV2 can significantly enhance feature extraction quality and reduce training time.
2. Deeper and More Complex Architectures: Extending the current CNN with more convolutional layers, residual connections, or attention blocks (e.g., Squeeze-and-Excitation Networks, CBAM) could improve performance, especially in distinguishing visually similar food items.
3. Data Balancing and Augmentation: Food-101 contains class imbalance; thus, techniques such as SMOTE or class-weighted loss functions should be implemented to improve performance on underrepresented categories.
4. Model Evaluation Across Devices: Future implementations should consider deploying the model on mobile or edge devices to measure latency, inference time, and user experience—key factors in real-world adoption.
5. Integration with Real-Time Systems: For practical applications, the model can be integrated with augmented reality (AR) food tracking, restaurant ordering systems, or nutritionist dashboards to enable real-time feedback.
6. Explainability and Interpretability: Incorporating explainable AI (XAI) techniques such as Grad-CAM can provide visual explanations for predictions, which are important in domains like health or personalized nutrition.
7. Multi-Modal Learning: Combining image data with textual descriptions (menus, ingredients) using multi-modal deep learning (e.g., CNN + Transformer) could further improve recognition accuracy..

CONCLUSION

Study has successfully demonstrated the implementation of a Convolutional Neural Network (CNN) for automatic food label detection using the Food-101 dataset. Despite being trained from scratch over a single epoch, the model achieved a baseline validation accuracy of 24.45%, revealing its capability to extract and learn basic visual patterns from complex food images. While the model's performance remains modest compared to state-of-the-art methods, it establishes a foundational framework for further development in food recognition systems. The experimental results and analysis confirm several important findings: (1) CNNs are capable of detecting coarse-grained food categories; (2) inter-class similarity and intra-class variability remain major challenges; and (3) training depth and architectural design significantly influence accuracy and generalization. To improve performance, future work should explore the integration of transfer learning, deeper CNN architectures, and advanced techniques such as attention mechanisms and data balancing strategies. These enhancements are essential to enable scalable deployment in real-world applications such as dietary monitoring, mobile health systems, and intelligent food service technologies. In conclusion, this research contributes valuable insights into image-based food classification and sets a solid groundwork for developing more accurate and robust food recognition systems using deep learning

REFERENCES

[1]. C. Bossard, M. Guillaumin, and L. Van Gool, “Food-101 – Mining Discriminative Components with Random Forests,” in Proc. European Conf. on Computer Vision (ECCV), Zurich, Switzerland, 2014, pp. 446–461.

[2]. D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in Proc. Int. Conf. on Learning Representations (ICLR), San Diego, CA, USA, 2015.

[3]. S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in Proc. ICML, 2015, pp. 448–456.

[4]. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv preprint arXiv:1409.1556, 2014.

[5]. A. Shorten and T. M. Khoshgoftaar, “A Survey on Image Data Augmentation for Deep Learning,” J. Big Data, vol. 6, no. 1, p. 60, 2019.

[6]. B. Yanto, E. Rouza, L. Fimawahib, B. H. Hayadi, and R. R. Pratama, “Penerapan Algoritma Deep

Learning Convolutional Neural Network dalam Menentukan Kematangan Buah Jeruk Manis Berdasarkan Citra Red Green Blue (RGB)," J. Teknol. Inform. dan Ilmu Komputer, vol. 10, no. 1, pp. 59–66, 2023.

[7]. M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. ICML, 2019, pp. 6105–6114.

[8]. B. Yanto, A. Lubis, B. H. Hayadi, and N. S. T. E. Armita, "Klarifikasi Kematangan Buah Nanas dengan Ruang Warna Hue Saturation Intensity (HSI)," INOVTEK Polbeng Seri Informatika, vol. 6, no. 1, pp. 135–146, 2021.

[9]. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. CVPR, Las Vegas, NV, USA, 2016, pp. 779–788.

[10]. Y. Zhang, X. Wang, and H. Ma, "Deep Learning in Food Category Recognition," Information Fusion, vol. 98, p. 101859, May 2023.

[11]. B. Yanto, M. A. Kartawidjaja, R. Sukwadi, and M. Bachtiar, "Implementation of Hue Saturation Intensity (HSI) Color Space Transformation Algorithm with Red, Green, Blue (RGB) Color Brightness in Assessing Tomato Fruit Maturity," Riau J. Comput. Sci., vol. 9, no. 2, pp. 167–178, 2023.

[12]. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proc. CVPR, 2017, pp. 2261–2269.

[13]. H. Meyers et al., "Food Recognition Using Mobile Devices: A Review," IEEE Access, vol. 9, pp. 21532–21549, 2021.

[14]. B. Yanto, B. Basorudin, J. Jufri, B. H. Hayadi, and N. S. T. E. Armita, "Indentifikasi Pola Aksara Arab Melayu Dengan Jaringan Syaraf Tiruan Convolutional Neural Network (CNN)," J. Scientific and Applied Informatics, vol. 3, no. 3, pp. 106–114, 2020.

[15]. Y. Liu, L. Zhang, and X. Wang, "Fine-Grained Food Recognition with Multi-Scale Attention," IEEE Access, vol. 8, pp. 85804–85815, 2020.

[16]. W. Liu et al., "SSD: Single Shot Multibox Detector," in Proc. ECCV, Amsterdam, Netherlands, 2016, pp. 21–37.

[17]. B. Yanto, L. Fimawahib, A. Supriyanto, B. H. Hayadi, and R. R. Pratama, "Klasifikasi Tekstur Kematangan Buah Jeruk Manis Berdasarkan Tingkat Kecerahan Warna dengan Metode Deep Learning Convolutional Neural Network," J. Inovtek Polbeng Seri Informatika, vol. 6, no. 2, pp. 259–268, 2021.

[18]. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, Cambridge, MA, USA: MIT Press, 2016.

[19]. T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006.

[20]. R. E. Pawening et al., "Food Object Detection Using Mask R-CNN for Diet Recommendation," J. Ilm. Tek. Inform. dan Komun., vol. 4, no. 1, pp. 87–97, 2023..