

# Hybrid convolutional vision transformer for extrusion-based 3D food-printing defect classification

Cholid Mawardi<sup>1,2</sup>, Agus Buono<sup>1</sup>, Karlisa Priandana<sup>1</sup>, Herianto<sup>3</sup>

<sup>1</sup>Computer Science Study Program, School of Data Science, Mathematics, and Informatics, Institut Pertanian Bogor, Bogor, Indonesia

<sup>2</sup>Department of Graphics Engineering Faculty of Industrial Technology, Politeknik Negeri Media Kreatif, Jakarta, Indonesia

<sup>3</sup>Department of Mechanical and Industrial Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia

## Article Info

### Article history:

Received Dec 12, 2024

Revised Jun 11, 2025

Accepted Jul 10, 2025

### Keywords:

3D food printing

Convolutional neural network

Hybrid convolutional

Image classification

Vision transformer

## ABSTRACT

Deep learning is generally used to perform remote monitoring of three-dimensional (3D) printing results, including extrusion-based 3D food printing. One of the widely used deep learning algorithms for defect detection in 3D printing is the convolutional neural network (CNN). However, the process requires high computational costs and a large dataset. This research proposes the Con4ViT model, a hybrid model that combines the strengths of vision transformer with the inherent feature extraction capabilities of CNN. The locally extracted features in the CNN were merged using the transformers' global features with four transformer encoder blocks. The proposed model has a smaller number of parameters compared to other lightweight pre-trained deep learning models such as VGG16, VGG19, EfficientNetB2, InceptionV3, and ResNet50. Thus, the proposed model is simplified. Simulations were conducted to classify defect and non-defect images obtained from the printing results of a developed extrusion-based 3D food printing device. Simulation results showed that the model produced an accuracy of 95.43%, higher than the state-of-the-art techniques, i.e., VGG16, VGG19, MobileNetV2, EfficientNetB2, InceptionV3, and ResNet50, with accuracies of 77.88%, 86.30%, 82.95%, 90.87%, 84.62%, and 93.83%, respectively. This research shows that the proposed Con4ViT model can be used for 3D food printing defect detection with high accuracy.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Karlisa Priandana

Computer Science Study Program, School of Data Science, Mathematics, and Informatics

Institut Pertanian Bogor

Bogor 16680, Indonesia

Email: karlisa@apps.ipb.ac.id

## 1. INTRODUCTION

Three-dimensional (3D) food printing technology is an innovation that enables the creation of foods with complex shapes and high precision using specialized 3D printers [1]. This technology works similarly to conventional 3D printers but uses food materials such as 'ink' to print food [2]. Food printing offers a range of advantages that enhance the culinary experience. It enables personalization, allowing for unique shapes, textures, and flavors according to individual preferences. In addition, since it uses precise techniques, it reduces food waste [3]. Food printing also encourages creativity in the kitchen, allowing new combinations of ingredients and designs that are impossible with conventional methods [4]. Luxury restaurants use 3D food printing to create unique dishes with artistic presentation [5], [6]. In the future, it has the potential to be used in the food industry for uniform and efficient mass production of food [7].

Defects in 3D food printing can affect the quality, appearance, and texture of the printed foods [8]. The causes of these defects can vary from technical issues with the printer and errors in printing parameters to the nature of the food material used [9]. Resolving these defects requires adjustments to the mold design, temperature, printing speed, and material settings [10]. Early detection of these defects can save the food materials used in 3D food printing by streamlining the process [11]. Early defect detection can be done remotely by taking an image of the food printing results obtained from a camera. Then, classification between defect and non-defect food images is done. A widely used method for defect and non-defect classification is deep learning convolutional neural networks (CNN) [12]. However, the CNN method requires high computing costs and large data.

Several models have been developed to reduce the computational cost of CNN, namely lightweight CNN models such as VGG16 [13], VGG19 [14], MobileNetV2 [15], EfficientNetB2 [16], InceptionV3 [17], and ResNet50 [18]. These algorithms use a parameter reduction approach to lower CNN computational costs. Another approach that has not been explored is simplifying the feature extraction process from large data. Vision transformer (ViT) is an algorithm that can perform global feature extraction from large amounts of data [19]. The ViT method has been proven to be able to classify tomography images for pulmonary nodule detection and diagnosis with good accuracy [20]. ViT offers several advantages over traditional CNN for computer vision tasks, including improved efficiency, scalability, transfer learning, performance, and flexibility [21]. With further research and development, ViT has the potential to become a powerful tool for a wide range of computer vision applications, such as crop pest image recognition [22].

In this research, we propose a hybrid model of CNN and ViT to combine the ability of local feature extraction in CNN with global feature extraction in ViT. The proposed method is called Con4ViT, which combines CNN with four transformer encoder blocks of ViT. Simulations were conducted to prove the performance of the proposed Con4ViT method for the developed extrusion-based 3D food printing device. Con4ViT is used to classify food printing images into two classes, namely defect and non-defect. Then, the proposed Con4ViT method is compared with the state-of-the-art techniques that have been mentioned, namely VGG16, VGG19, MobileNetV2, EfficientNetB2, InceptionV3, and ResNet50.

The rest of this paper is structured as follows. Section 2 describes the related works about deep learning models in 3D Printing. Section 3 presents the methodology of this research, including the dataset acquired from a developed extrusion-based 3D food printing and the proposed architecture of Con4ViT. Section 4 provides the results and discussion of the model performance, and section 5 presents the main conclusions of this work.

## 2. RELATED WORKS

Baumann and Roller [23] conducted early research on defect control in 3D printing machines. The study involves computer vision to detect fault diagnosis, dividing the defect classification into five classes, namely detachment, missing material flow, deformed object, surface errors, and deviation from the model. Three classes were successfully detected from the five classes, with a detection rate of 60 to 80% [23]. Rachmawati *et al.* [24], introduced data augmentation for 3D printing to vary the amount of data to help reduce overfitting. The study used a regular CNN, and the accuracy of the study was 95.45%. Other studies that utilize deep learning in 3D printing are summarized in Table 1. As seen in Table 1, previous research using the ResNet50 model with a 3D food printing image dataset of chocolate objects resulted in an accuracy of 93.80% [25]. The study used pre-trained InceptionV3 and ResNet 50 models with additional hyperparameter tuning on learning rate to obtain the optimum value. Then, the research conducted by Paraskevoudis *et al.* [26] monitored defects in fused fluid fabrication (FFF) 3D printing. The study used the VGG16 pre-trained architecture model as a base network with 16 convolutional layers and 3 fully connected layers. The resulting model accuracy is 92.70%.

Table 1. Performance comparison of different deep learning models in 3D printing

Works	Machine (material)	Method	Accuracy (%)	Sensitivity (%)	Precision (%)	F1-score (%)
Rachmawati <i>et al.</i> [24]	3D Printing	CNN+MobileNet	95.45	-	-	-
Baumgartl <i>et al.</i> [27]	3D Printing	CNN+Classic ML	96.80	96.80	96.52	96.42
						(kappa score)
Mawardi <i>et al.</i> [25]	3D food printing	ResNet50	93.80	96.56	96.84	96.70
Paraskevoudis <i>et al.</i> [26]	3D Printing	VGG16	92.70	92.00	75.01	82.10

Defect classification in 3D food printing generally exhibits lower accuracy compared to traditional 3D printing. This discrepancy arises from the differences in printing materials, which pose challenges for computer vision systems. While 3D food printing utilizes soft materials like chocolate and pasta, traditional 3D printing employs more rigid materials that are easier to analyze for object detection and image classification [28], [29]. Given these challenges, deep learning models are considered well-suited for defect detection in 3D food printing.

### 3. METHODOLOGY

This section outlines the process for detecting and classifying print results from 3D food printing devices into two categories: defect and non-defect. Classification is performed using a newly proposed algorithm a hybrid model that combines a CNN with a ViT on images captured from the 3D food printing device. The defect detection process is illustrated in Figure 1.

The first stage involves data collection, where videos of the food being printed are recorded using an Ender-V3 3D printer equipped with a Luckybot extruder. Video capture is facilitated by OctoPrint plugins. These videos are then segmented into individual image frames, which are manually labeled as either defect or non-defect based on the actual condition of the printed results. Then, data preprocessing is conducted on the labeled images. The dataset is then split into 80% training data and 20% validation data. The next step involves developing the hybrid model, which integrates CNN and ViT components through several transformer encoder blocks. During the training phase, validation is performed to mitigate the risk of overfitting. To assess the model's performance, a confusion matrix is employed.

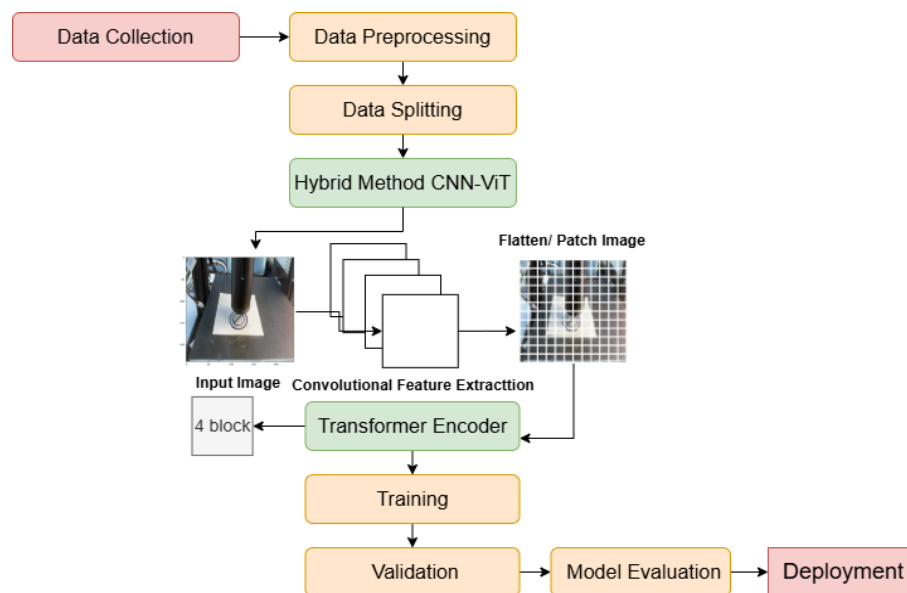


Figure 1. The research methodology

#### 3.1. Data collection

The initial process for capturing images of 3D food printing involves using a Logitech C270 webcam to record the printing procedure, as illustrated in Figure 2. A Raspberry Pi microcontroller serves as the interface between the 3D food printing device and the computer, enabling the detection and recording of the printing process. Various printing tasks are conducted using different designs, from which two outcomes are selected: one representing a defect and the other a non-defect.

After the printing is completed, the recordings are then segmented into individual images. For the defect category, which includes the failed print video with a duration of 2 minutes and 11 seconds, images are extracted every second, resulting in a total of 262 images. Similarly, the non-defect category, which has a duration of 2 minutes and 12.5 seconds, produces 265 images. The images from the defect process are categorized as defect samples, while those from the non-defect process are classified as non-defect samples. In addition, the dataset is supplemented with images from a regular 3D printing device. This inclusion adds variety to the dataset and enhances data representation, ultimately improving the accuracy of the model.

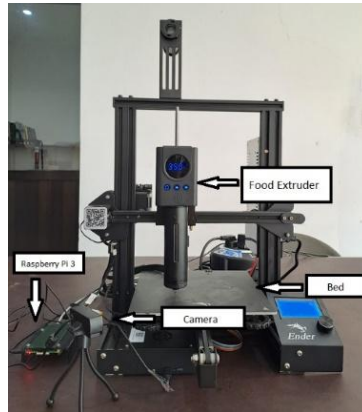


Figure 2. Experimental setup for defect detection in 3D food printing

### 3.2. Data preprocessing

This section describes the data preprocessing steps necessary to prepare the images for effective processing by the deep learning model. The preprocessing techniques include resizing, rescaling, and data augmentation. Initially, the 3D food printing images are captured from the camera and resized to  $128 \times 128$  pixels [30]. Following this, image scaling is applied to adjust the pixel values from the range of  $[0, 255]$  to  $[0, 1]$  [31]. This rescaling is crucial for preventing pixel values from becoming excessively large or small, which can lead to numerical instability and slow down the computational process [32]. All experiments are conducted using the Keras library in Python, utilizing an A100 GPU with 150 GB of memory.

In this research, various data augmentation techniques are employed to enhance the dataset and facilitate the hybrid modeling process between CNN and ViTs. These techniques are designed to mitigate overfitting and improve the overall accuracy of the model. The augmentation methods used include width shift, height shift, zoom range, flip, and rotation range [33]. A complete summary of the augmentation techniques applied to the 3D food printing image dataset is presented in Table 2.

Table 2. Values and parameters of the applied transformation techniques

Parameters	Value of parameters	Action
Width shift range	0.2	Randomly adjusts the image's horizontal size by 20%.
Height shift range	0.2	Randomly adjusts the image's vertical size by 20%.
zoom_range	0.2	Extend the zoom by 0.2 from the center.
shear_range	0.2	0.2 is the image's extension.
rotation_range	10	Spin in a -10 to a-10-degree circle.
rescale	1./255	scales (normalizes) the image pixel values to fall within the range of 0 to 1, from an initial value range of 0 to 255.

### 3.3. Data splitting

The dataset is divided in an 80:20 ratio, with 80% allocated for training and 20% reserved for validation. This split is consistently applied to both the Con4ViT model and other benchmark models to ensure that the results are comparable. By maintaining the same training and validation data distribution across all models, we can confidently attribute any observed differences in performance to variations in model architecture rather than inconsistencies in the data. This approach enhances the reliability of the evaluation and strengthens the conclusions drawn from the comparative analysis.

### 3.4. Hybrid method CNN-ViT (Con4ViT)

This section explains the functionality of the Con4ViT model, which combines the strengths of CNNs and ViTs to effectively capture both local and global features in images. The model begins with local feature extraction through a convolutional block comprising three layers. After the convolutional operations are performed, the resulting multi-dimensional output is flattened into a one-dimensional vector. This vector is then processed by the transformer encoder, which utilizes a self-attention mechanism to recognize the relationship between elements in the vector across four transformer encoder blocks. A complete block diagram illustrating the architecture of the proposed Con4ViT hybrid model is shown in Figure 3.

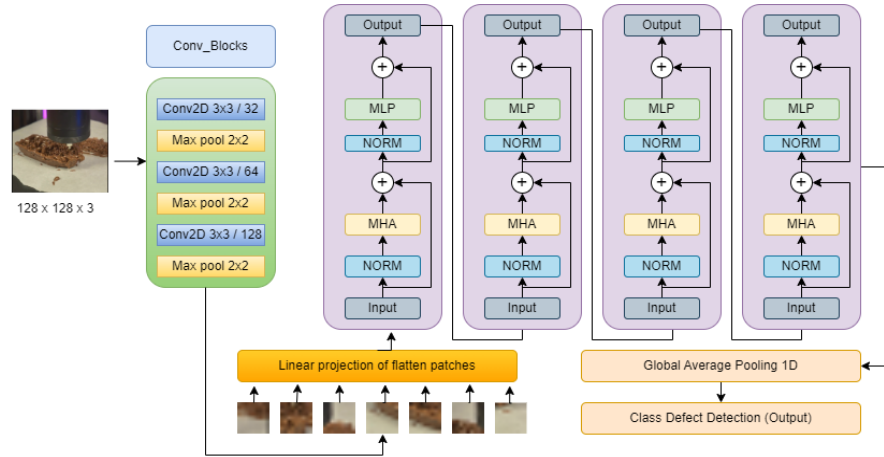


Figure 3. The proposed Con4ViT hybrid method

The input image of size  $128 \times 128 \times 3$  is fed into CNN to extract local features [34] with sequential CNNs consisting of 3 convolutional and max-pooling layers. The convolution layer utilized rectified linear unit (ReLU) activation function as shown in (1) [34].

$$y(i, j) = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} (i + m, j + n) \cdot w(m, n) \quad (1)$$

Where  $x(i, j)$  is the image input in pixel  $(i, j)$ ,  $w(m, n)$  is the weight of kernel/filter with size  $M \times N$  and  $y(i, j)$  is the output after the convolution operation at position  $(i, j)$ . Then, the pooling layer utilizes the (2).

$$z(p, q) = \max(\{y(2p + m, 2q + n) \mid m, n \in \{0, 1\}\}) \quad (2)$$

Where  $z(p, q)$  is the output after the max pooling operation at position  $(p, q)$ , the indices  $m$  and  $n$  iterate over the  $2 \times 2$  pooling window, and the stride  $s$  is 2, indicating that the pooling window moves 2 pixels at a time in both dimensions. The pooling operation reduces the input dimension by taking the maximum value of each sub-area in the input matrix. If the pooling size is  $2 \times 2$ , from each  $2 \times 2$  block, the maximum value is taken as the pooling result.

After the pooling operation, a combination with flatten is performed using the reshape feature with the encoder standard. In flatten, the image is processed into patches so that it can be converted into a vector sequence. As seen in Figure 4, the 3D food printing input image is processed into non-overlapping patches. In this process, the original image in Figure 4(a) is first divided into multiple smaller regions in Figure 4(b), each of size  $20 \times 20$  pixels. These patches are then transformed into one-dimensional vectors through a flatten operation. Flatten converts a multi-dimensional tensor into a one-dimensional vector without changing the values of the elements in the tensor. For example, if the input is a 3D tensor with size  $(batch\_size, h, w, c)$  (e.g., from the convolution layer), then flatten will convert it into a 2D tensor of size  $(batch\_size, h \times w \times c)$ . Mathematically, for an input  $x_{1,1,1}, x_{1,1,2} \dots$  of size  $(h, w, c)$  the result is (3).

$$Flatten(X) = [x_{1,1,1}, x_{1,1,2} \dots, x_{h,w,c}] \quad (3)$$

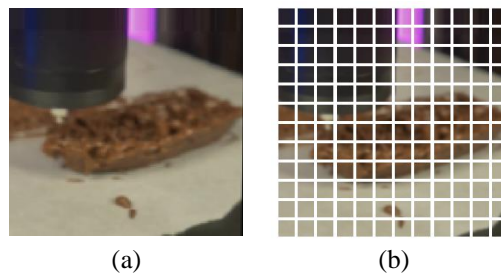


Figure 4. Patch of flatten image 3D food printing of (a) input image and (b) patch image

Then, the image patches with the flatten process enter the encoder transformer, with a layer normalization process for multi-head attention. Multi-head attention in the transformer model calculates the attention weight in (4), as explained in [19].

$$Attention(Q, K, V) = softmax\left(\frac{QK^t}{\sqrt{d_k}}\right)v \quad (4)$$

A SoftMax function converts these attention values into a multi-head attention probability distribution. It also allows the model to focus on the input's more important or relevant parts based on the  $Q$  and  $K$  values and assign measured values to the selected information. After the attention process, it goes to the feed-forward network (FFN), which is a linear transformation operation [19], as shown in (5):

$$FFN(x) = \max(0, xW_1 + b_1) W_2 + b_2 \quad (5)$$

Where  $x$  is the input to the FFN,  $W_1$ , and  $W_2$  are weight matrices,  $b_1$  and  $b_2$  are bias vectors. The operation involves first applying a linear transformation, followed by a ReLU activation function (represented by  $\max(0, \cdot)$ ), and then applying another linear transformation. The operation  $xW_1 + b_1$  is a linear operation in the first layer and the hidden layer, and then the result passes through the ReLU activation function. The result of the activation function is passed to the next layer, where it is multiplied by the weights  $W_2$  and added with the bias  $b_2$  to give the final output. The next step is to calculate the loss function, described in the (6).

$$L = -\frac{1}{N} \sum_{i=1}^N \text{Log}(p_{true}) \quad (6)$$

Where  $L$  is the overall loss value for the batch of predictions,  $N$  is the number of samples, and  $p_{true}$  is the probability of the correct class with a loss function calculating how significant the difference is between the model's predicted probability and the actual label.

### 3.5. Model training and validation

The next step involves training the model using the dataset for a total of 30 epochs, during which model parameters are adjusted to enhance performance. Validation data is utilized to assess the model's effectiveness throughout this process. The Adam optimizer is employed to optimize the model, ensuring efficient convergence during training. Training is conducted multiple times to cover all architectures being compared, including the proposed Con4ViT model, VGG16, VGG19, MobileNetV2, EfficientNetB2, InceptionV3, and ResNet50. This comprehensive approach allows for a thorough evaluation of each model's performance.

### 3.6. Model evaluation and performance evaluation

In this study, the performance of the Con4ViT model for 3D food printing defect classification is evaluated using a confusion matrix [35]. This matrix summarizes the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These four categories enable the calculation of key performance metrics: accuracy, recall, precision, and F1-score, defined by (7)-(10) [36].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

$$F1 - score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (10)$$

Additionally, the gradient-weighted class activation mapping (Grad-CAM) method will be used to analyze the image regions that are crucial for determining classification results [37]. Grad-CAM is a visualization technique in deep learning that highlights important areas of an image that influence the model's predictions. It generates a heatmap indicating the significant regions for the predicted class, calculated using the (11).



$$L^c = \text{ReLU}(\sum_k \alpha_k^c A^k) \quad (11)$$

Here, the weights are computed through global average pooling of the gradients as in (12).

$$A^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (12)$$

Where  $A^k$  represents the activation from the  $k^{\text{th}}$  filter in the last layer. Once the heatmap is generated, it is reshaped to  $28 \times 28$  pixels and overlaid onto the original image, with color coding to highlight the important areas. Grad-CAM provides valuable visual insights into the regions that the model focuses on, enhancing interpretability and understanding of the model's decision-making process.

#### 4. RESULTS AND DISCUSSION

This section presents the results of data collection, data preprocessing, Grad-CAM analysis, and experiments for the performance evaluation of the proposed and developed Con4ViT model for defect and non-defect classification in 3D food printing. The comparative performance of the proposed model with other pre-trained based models is also explained in this section.

##### 4.1. Data collection

As a result of the data collection stage, we obtained 2,085 images as a combination of 527 print results image from a 3D food printing device and 1,558 print results from a 3D printing device. Based on the 80:20 ratio, the training data consists of 1,669 images, and the validation data consists of 416 images. Figure 5 shows examples from the 3D food printing dataset, divided into defect and non-defect categories.

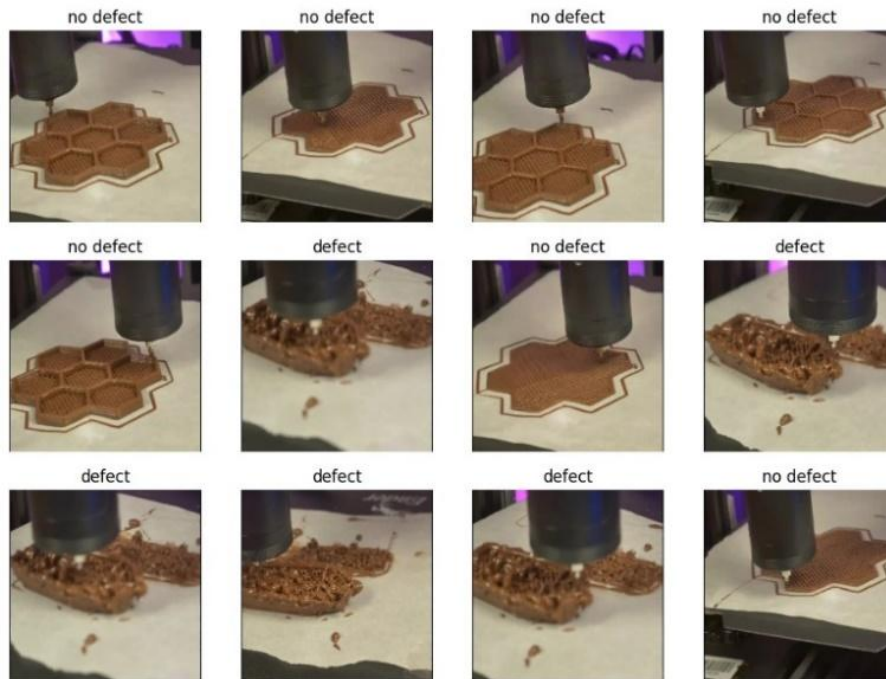


Figure 5. Image dataset example from the 3D food printing device, utilizing chocolate as the material

##### 4.2. Data preprocessing

The technique of data preprocessing in the form of data augmentation that produces images as seen in Figure 6. Figure 6(a) shows the original image of 3D food printing, Figure 6(b) is a rotation with a value of 10% from the initial position, Figure 6(c) enlarges the display with zoom\_range from a scale of 20%. In Figure 6(d), width\_share\_range is also done by shifting the image by 20%, and in Figure 6(e), the image height adjusts to the height shift range with 20% of the original image.

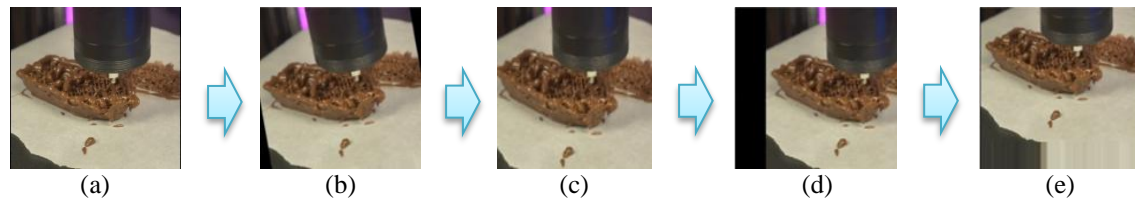


Figure 6. Result data preprocessing 3D food printing of (a) original image, (b) rotation, (c) zoom\_range, (d) width\_share\_range, and (e) height\_share\_range

#### 4.4. Training and validation

In Figure 7, training and validation were performed on the Con4ViT model with 30 epochs. The model training process is seen in the blue line, while the model validation uses the red line. The results obtained are the accuracy results in training of 98.20% and the accuracy results in validation of 95.91%.

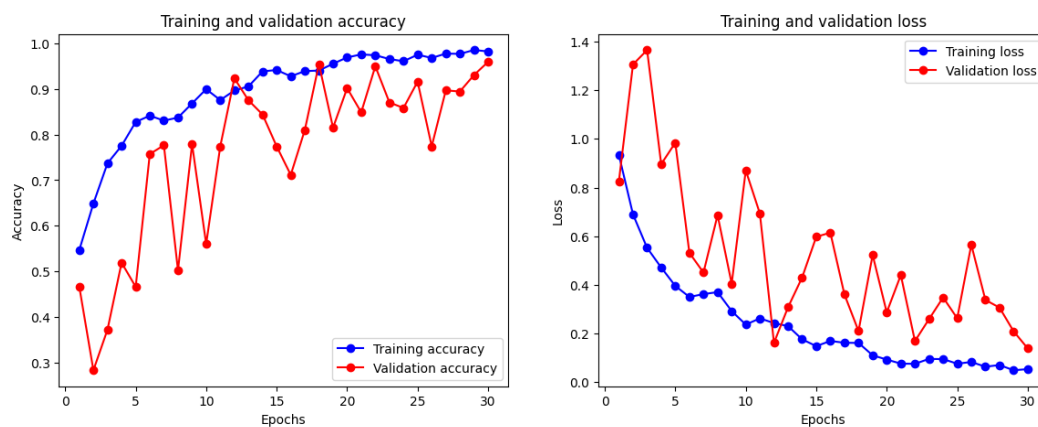


Figure 7. Training and validation Con4ViT model

#### 4.3. Model evaluation

Figure 8 shows the confusion matrix of the Con4ViT model. In Figure 8(a), it can be seen that the proposed model with 416 validation data has good performance, with 199 images correctly classified as defect (TP) and 200 images correctly classified as non-defect (TN). 9 images correctly classified as defect (FP), 8 images correctly classified as non-defect (FN). In Figure 8(b), it can also be seen that the model used when using the entire data, namely 2,085 images, with 1,024 images correctly classified as defect (TP) and 1,010 images correctly classified as non-defect (TN). 13 images correctly classified as defect (FP) 30 images correctly classified as non-defect (FN). With the results of the Con4ViT model evaluation performance using data validation, good results were obtained, namely, accuracy, precision, recall, and F1-score. The accuracy of the Con4ViT model reached 95.91%, with a precision of 95.69%, a sensitivity of 96.15%, and an F1-score of 95.92%.

#### 4.4. Grad-CAM analysis

This model was performed with additional analysis using visualization to visually understand which parts of the image are considered necessary and contribute to the model's predictions [37]. Figure 9 shows the heatmap visualization area, which is the critical area focused on by the 3D food printing image. The visual focus is close to the lighter or blue boundary of the heatmap, which shows the surrounding area that has the most significant influence on the model prediction. In applying Grad-CAM to the Con4ViT model, Figure 9(a) shows the orange and red colors on the edge of the design and slightly below the nozzle of the 3D food printing extruder. Figure 9(b) shows the red and orange areas around the print under the nozzle of the printing head, and the blue color is at the nozzle point. Figure 9(c) covers more surfaces around the print area, with the hot color spread over a wider area, while the blue color is in the inner part of the print process. Overall, Grad-CAM can recognize relevant visual features to identify or monitor print activity and indicate essential parts of the image for the predicted class.



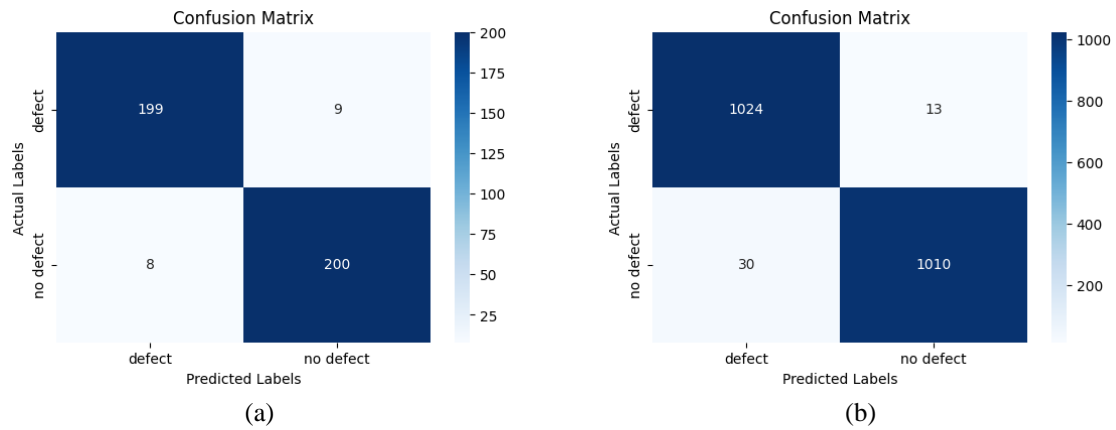


Figure 8. Predicted label with confusion matrix of (a) model evaluation with data validation and (b) model evaluation with all data

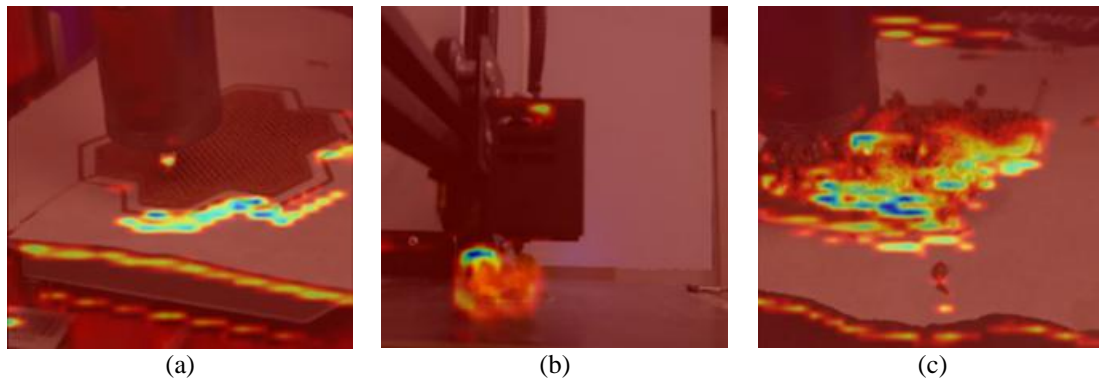


Figure 9. Grad-CAM image of the 3D food printing of (a) Grad-CAM non defect, (b) Grad-CAM defect with nozzle focus, and (c) Grad-CAM defect with wider area

#### 4.5. Comparison of Con4ViT model with another pre-trained model

To further evaluate the model's performance, the proposed Con4ViT model was compared with other CNN models based on pre-trained learning, namely VGG16, VGG19, MobileNetV2, EfficientNetB2, InceptionV3, and ResNet50. Table 3 compares our proposed Con4ViT model performance with other pre-trained deep-learning models. The resulting performance results were 95.91% accuracy, 95.69% precision, 96.15% recall, and 95.92% F1-score.

Table 3. Comparison of the Con4ViT model approach with other pre-trained models

Model	Parameter (million)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
VGG16	17.9	77.88	85.89	66.80	75.15
VGG19	23.2	86.30	86.10	86.83	86.46
MobileNetV2	2.4	82.95	87.28	77.29	81.98
Con4ViT	6.7	95.91	95.69	96.15	95.92
EfficientNetB2	9.3	90.87	90.76	91.11	90.93
InceptionV3	22.3	84.62	91.24	90.51	90.97
ResNet50	23.8	93.83	96.84	96.56	96.70

One of the key findings in this comparison is that the Con4ViT model has a relatively low parameter count of 6.7 million, especially when compared to larger models such as VGG19 (23.2 million) and ResNet50 (23.8 million). This smaller parameter count indicates that Con4ViT is more lightweight, making it a good choice for deployment in resource-constrained environments with limited computing power. Despite having fewer parameters, Con4ViT achieved the highest accuracy of 95.91%, far outperforming all other models listed in correctly predicting outcomes on the evaluation dataset.

When looking at precision, recall, and F1-score, Con4ViT consistently leads in all these metrics. It has an impressive precision of 95.69%, meaning that when it predicts the positive class, it is more likely to be correct, which is essential in applications where negative positives are detrimental. Its recall score is also high at 96.15%, indicating the model's ability to identify a large proportion of actual positive cases accurately. With an F1-score of 95.92%, Con4ViT stands out as the best-performing model in terms of overall balanced performance.

In comparison, VGG16 and VGG19 have higher parameter counts but lower performance metrics, particularly in recall and F1-scores, indicating that they struggle to balance accuracy and efficiency. MobileNetV2, while lightweight with only 2.4 million parameters, does not achieve the same level of performance as Con4ViT across all metrics. EfficientNetB2 and InceptionV3 deliver competitive results, but both must catch up to Con4ViT's metrics. While EfficientNetB2 has a moderate parameter count (9.3 million) and solid accuracy, more is needed to achieve the overall performance level of Con4ViT, indicating that simply being efficient in terms of parameters does not guarantee better results. ResNet50 achieves high metrics, particularly in the F1-score (96.70%), but does not outperform Con4ViT in any individual metrics and has a much larger parameter count.

In conclusion, this analysis shows that the Con4ViT model outperforms all other comparison models in terms of accuracy, precision, recall, and F1-score. This makes it an excellent choice for tasks that require high accuracy and model efficiency. Its lower parameter count and excellent performance metrics suggest that this model can be very effective for a wide range of applications, especially where computational resources are a constraint.

The results shown in Figure 10 showed that the training and validation performance of Con4ViT on 3D food printing defect classification has low fluctuation. However, with few parameters, the final performance value on Con4ViT has good results. EfficientNetB2 is better at maintaining stable validation accuracy by showing better generalization. Overall, the tested pre-training models have good values, but the proposed Con4ViT model has good accuracy results so that it can be used in other research sets, such as large or small data sets.

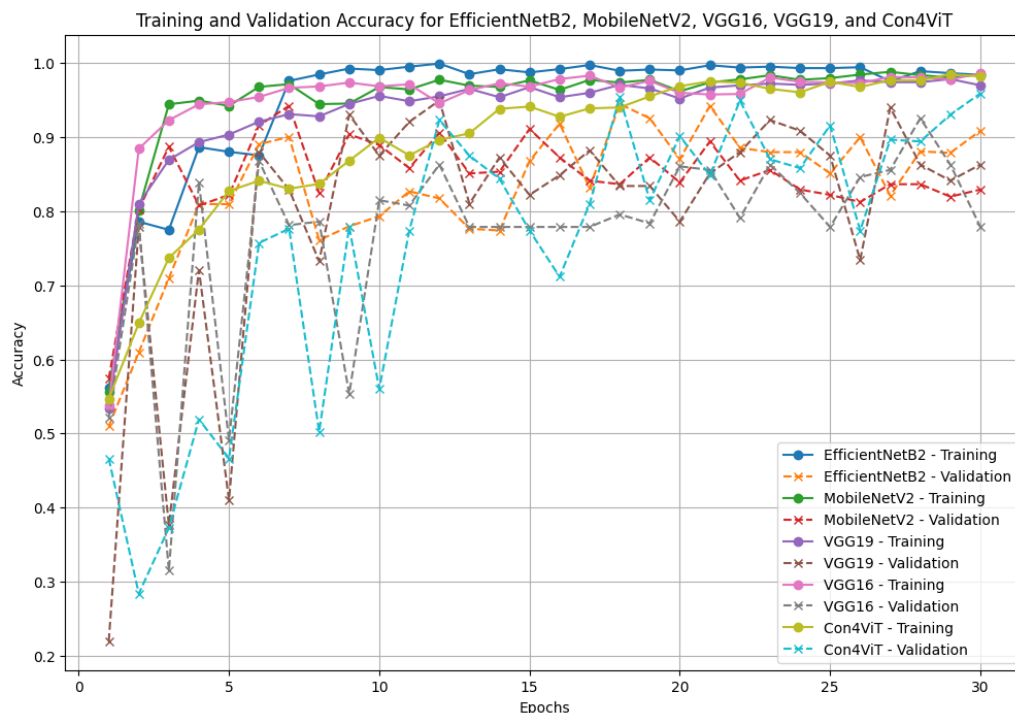


Figure 10. Training and validation performance compared to some other methods with Con4ViT

## 5. CONCLUSION

This paper proposes a hybrid method combining CNN with ViT on 3D food printing defect classification. For this purpose, we conducted experiments with 2,085 data from the 3D food printing

process. In the experiment, images obtained from a 3D food printing device are divided into two classes, namely defect and non-defect classes. Image preprocessing, including resizing and rescaling data augmentation, is very influential in this research. Then, a Con4ViT model is built with a combination of CNN and ViT features where multiple CNN layers extract local features, and the ViT model captures context on global features with a 4-block transformer encoder using a self-attention mechanism. Pre-trained models, including VGG16, VGG19, MobileNetV2, EfficientNetB2, InceptionV3, and ResNet50, are compared as a performance comparison. Con4ViT has good performance of defect classification on 3D food printing images compared to other pre-trained with 95.91% accuracy. The experimental results have few parameters and low computation, which will be easier to implement on IoT devices for smartphone-based defect monitoring in the future.

ACKNOWLEDGEMENTS

The authors would like to express their appreciation to the Directorate of Research and Innovation at IPB University for supporting the management of this grant.

FUNDING INFORMATION

This research was funded by the Directorate General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia under the Doctoral Dissertation Grant scheme (*Penelitian Disertasi Doktor-PDD*), grant number 027/E5/PG.02.00.PL/2024 jo. 22105/IT3.D10/PT. 01.03/P/B/2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Cholid Mawardi	✓	✓	✓	✓	✓			✓	✓	✓	✓		✓	✓
Agus Buono	✓	✓				✓	✓			✓		✓		✓
Karlisa Priandana	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓
Herianto	✓	✓				✓	✓			✓		✓		✓

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the first author, [CM] or corresponding author, [KP]. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

REFERENCES

[1] N. Nachal, J. A. Moses, P. Karthik, and C. Anandharamakrishnan, "Applications of 3D printing in food processing," *Food Engineering Reviews*, vol. 11, no. 3, pp. 123–141, 2019, doi: 10.1007/s12393-019-09199-8.

[2] J. Lee, "A 3D food printing process for the new normal era: a review," *Processes*, vol. 9, no. 9, 2021, doi: 10.3390/pr9091495.




[3] R. Soni, K. Ponappa, and P. Tandon, "A review on customized food fabrication process using food layered manufacturing," *Lwt*, vol. 161, no. October 2021, 2022, doi: 10.1016/j.lwt.2022.113411.

[4] M. A. Augustin, C. J. Hartley, G. Maloney, and S. Tyndall, "Innovation in precision fermentation for food ingredients," *Critical Reviews in Food Science and Nutrition*, vol. 64, no. 18, pp. 6218–6238, 2024, doi: 10.1080/10408398.2023.2166014.




[5] I. Shabir *et al.*, "Advancements in food printing technologies and their potential culinary applications: a contemporary exploration," *Journal of Food Processing and Preservation*, vol. 2024, 2024, doi: 10.1155/2024/6621344.

- [6] M. G. J. Meijers and D. I. Han, "The 3D food printing pyramid of gastronomy: a structured approach towards a future research agenda," *International Journal of Gastronomy and Food Science*, vol. 37, no. April, 2024, doi: 10.1016/j.ijgfs.2024.100969.
- [7] M. Waseem, A. U. Tahir, and Y. Majeed, "Printing the future of food: the physics perspective on 3D food printing," *Food Physics*, vol. 1, no. July 2023, p. 100003, 2024, doi: 10.1016/j.foodp.2023.100003.
- [8] T. Pereira, S. Barroso, and M. M. Gil, "Food texture design by 3d printing: a review," *Foods*, vol. 10, no. 2, pp. 1–26, 2021, doi: 10.3390/foods10020320.
- [9] A. O. Agunbiade *et al.*, "Potentials of 3D extrusion-based printing in resolving food processing challenges: a perspective review," *Journal of Food Process Engineering*, vol. 45, no. 4, pp. 1–31, 2022, doi: 10.1111/jfpe.13996.
- [10] T. Sivarupan *et al.*, "A review on the progress and challenges of binder jet 3D printing of sand moulds for advanced casting," *Additive Manufacturing*, vol. 40, 2021, doi: 10.1016/j.addma.2021.101889.
- [11] W. L. Ng, G. L. Goh, G. D. Goh, J. S. J. Ten, and W. Y. Yeong, "Progress and opportunities for machine learning in materials and processes of additive manufacturing," *Advanced Materials*, vol. 36, no. 34, 2024, doi: 10.1002/adma.202310006.
- [12] L. Zhu, P. Spachos, E. Pensini, and K. N. Plataniotis, "Deep learning and machine vision for food processing: a survey," *Current Research in Food Science*, vol. 4, no. December 2020, pp. 233–249, 2021, doi: 10.1016/j.crfs.2021.03.009.
- [13] H. Qassim, D. Feinzimer, and A. Verma, "Residual squeeze VGG16," *arXiv-Computer Science*, pp. 1–11, 2017.
- [14] T. Carvalho, E. R. S. De Rezende, M. T. P. Alves, F. K. C. Balieiro, and R. B. Sovat, "Exposing computer generated images by eye's region classification via transfer learning of VGG19 CNN," *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Cancun, Mexico, 2017, pp. 866–870, doi: 10.1109/ICMLA.2017.00-47.
- [15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.
- [16] S. Kalvankar, H. Pandit, and P. Parwate, "Galaxy morphology classification using efficientNet architectures," *arXiv-Computer Science*, pp. 1–13, 2021.
- [17] C. Wang *et al.*, "Pulmonary image classification based on inception-v3 transfer learning model," *IEEE Access*, vol. 7, pp. 146533–146541, 2019, doi: 10.1109/ACCESS.2019.2946000.
- [18] I. Z. Mukti and D. Biswas, "Transfer Learning based plant diseases detection using ResNet50," *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, Khulna, Bangladesh, 2019, pp. 1–6, doi: 10.1109/EICT48899.2019.9068805.
- [19] A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global context vision transformers," *Proceedings of Machine Learning Research*, vol. 202, pp. 12633–12646, 2023.
- [20] R. Li, C. Xiao, Y. Huang, H. Hassan, and B. Huang, "Deep learning applications in computed tomography images for pulmonary nodule detection and diagnosis: a review," *Diagnostics*, vol. 12, no. 2, 2022, doi: 10.3390/diagnostics12020298.
- [21] J. Mauricio, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: a literature review," *Applied Sciences*, vol. 13, no. 9, 2023, doi: 10.3390/app13095521.
- [22] X. Fu *et al.*, "Crop pest image recognition based on the improved ViT method," *Information Processing in Agriculture*, vol. 11, no. 2, pp. 249–259, 2024, doi: 10.1016/j.inpa.2023.02.007.
- [23] F. Baumann and D. Roller, "Vision based error detection for 3D printing processes," *MATEC Web of Conferences*, vol. 59, pp. 3–9, 2016, doi: 10.1051/mateconf/20165906003.
- [24] S. M. Rachmawati, M. A. Paramartha Putra, T. Jun, D. S. Kim, and J. M. Lee, "Fine-tuned CNN with data augmentation for 3D printer fault detection," *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, Republic of, 2022, pp. 902–905, doi: 10.1109/ICTC55196.2022.9952484.
- [25] C. Mawardi, A. Buono, K. Priandana, and H. Herianto, "Performance analysis of ResNet50 and inception-V3 image classification for defect detection in 3D food printing," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 14, no. 2, pp. 798–804, 2024, doi: 10.18517/ijaseit.14.2.19863.
- [26] K. Paraskevoudis, P. Karayannis, and E. P. Koumoulos, "Real-time 3D printing remote defect detection (stringing) with computer vision and artificial intelligence," *Processes*, vol. 8, no. 11, pp. 1–15, 2020, doi: 10.3390/pr8111464.
- [27] H. Baumgartl, J. Tomas, R. Buettner, and M. Merkel, "A deep learning-based model for defect detection in laser-powder bed fusion using in-situ thermographic monitoring," *Progress in Additive Manufacturing*, vol. 5, no. 3, pp. 277–285, 2020, doi: 10.1007/s40964-019-00108-3.
- [28] K. Prabha *et al.*, "Recent development, challenges, and prospects of extrusion technology," *Future Foods*, vol. 3, Jun. 2021, doi: 10.1016/j.fufo.2021.100019.
- [29] M. Shahbazi and H. Jäger, "Current status in the utilization of biobased polymers for 3D printing process: a systematic review of the materials, processes, and challenges," *ACS Applied Bio Materials*, vol. 4, no. 1, pp. 325–369, 2021, doi: 10.1021/acsabm.0c01379.
- [30] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The impact of pre- and post-image processing techniques on deep learning frameworks: a comprehensive review for digital pathology image analysis," *Computers in Biology and Medicine*, vol. 128, 2021, doi: 10.1016/j.combiomed.2020.104129.
- [31] G. Ghiasi, X. Gu, Y. Cui, and T. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *Computer Vision – ECCV 2022*, Cham, Springer, 2022, pp. 540–557. doi: 10.1007/978-3-031-20059-5\_31.
- [32] N. Mohammad, A. M. Muad, R. Ahmad, and M. Y. P. M. Yusof, "Accuracy of advanced deep learning with TensorFlow and Keras for classifying teeth developmental stages in digital panoramic imaging," *BMC Medical Imaging*, vol. 22, no. 1, pp. 1–13, 2022, doi: 10.1186/s12880-022-00794-6.
- [33] K. Alomar, H. I. Aysel, and X. Cai, "Data augmentation in classification and segmentation: a survey and new strategies," *Journal of Imaging*, vol. 9, no. 2, 2023, doi: 10.3390/jimaging9020046.
- [34] K. S. R. Sekhar, T. R. Babu, G. Prathibha, K. Vijay, and L. C. Ming, "Dermoscopic image classification using CNN with handcrafted features," *Journal of King Saud University - Science*, vol. 33, no. 6, pp. 1–9, Sep. 2021, doi: 10.1016/j.jksus.2021.101550.
- [35] N. A. M. Roslan, N. M. Diah, Z. Ibrahim, Y. Munarko, and A. E. Minarno, "Automatic plant recognition using convolutional neural network on Malaysian medicinal herbs: the value of data augmentation," *International Journal of Advances in Intelligent Informatics*, vol. 9, no. 1, pp. 136–147, 2023, doi: 10.26555/ijain.v9i1.1076.
- [36] Q. Zhang, Q. Yang, X. Zhang, Q. Bao, J. Su, and X. Liu, "Waste image classification based on transfer learning and convolutional neural network," *Waste Management*, vol. 135, pp. 150–157, 2021, doi: 10.1016/j.wasman.2021.08.038.
- [37] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: why did you say that?," *arXiv-Statistics*, pp. 1–4, 2017.




**BIOGRAPHIES OF AUTHORS**

**Cholid Mawardi**    received a bachelor's degree from the Department of Information Systems STMIK Jakarta STI&K, Indonesia, in 2014, a master's degree from the Department of Electrical Engineering, Mercu Buana University, Indonesia, in 2018, and is currently pursuing a doctoral program in computer science at IPB University. He works at the Department of Graphics Engineering, Politeknik Negeri Media Kreatif, Indonesia. His research includes 3D printing optimisation, graphics engineering, information systems, computational intelligence, and deep learning. He can be contacted at email: mawardicholid@apps.ipb.ac.id.






**Agus Bueno**    is a Professor from Bogor Agricultural University (IPB), Bogor, Indonesia. Working as a lecturer at IPB University, he earned his Bachelor of Science in Statistics from IPB University, Bogor, Indonesia; Master's degree in Statistics from IPB University, Bogor; and Ph.D. in Computer Science from University of Indonesia, Indonesia. His fields of interest includes mathematics, computer science, data science, artificial intelligence, modeling and computing, new science, and pattern recognition. He is currently served as the Dean of the School of Data Science, Statistics, Mathematics, and Informatics at IPB University. He can be contacted via email at agusbuono@apps.ipb.ac.id.



**Karlisa Priandana**    (Senior Member, IEEE) is an Associate Professor at IPB University (Institut Pertanian Bogor), where she has been a lecturer of Computer Science since 2012. She earned her Bachelor's degree in Electrical Engineering from Bandung Institute of Technology, Indonesia; her Master's degree in International Development Engineering from Tokyo Institute of Technology, Japan; and her Doctoral degree in Electrical Engineering from Universitas Indonesia. Her research interests include robotics and artificial intelligence. She has received several awards and recognitions, including Best Student Award (Ganesha Prize) from Bandung Institute of Technology (2006) and Best Doctoral Graduate from Universitas Indonesia (2017). She was also awarded several prestigious scholarships, including: a full scholarship from the Bandoengse Technische Hoogeschool Fonds (BTHF) for a short-term research program at TU Delft, The Netherlands (2008); a Monbukagakusho (MEXT) scholarship for her Master study in Japan; a full doctoral and a full post-doctoral scholarships from the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia for her doctoral study in Universitas Indonesia (2014-2017) and her short research program in Tampere University, Finland (2021). In 2025, she was entrusted as the Acting Director of Talent Development for Research and Development at the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia. She can be contacted via email at karlisa@apps.ipb.ac.id.



**Herianto**    is a Professor of additive manufacturing systems at Universitas Gadjah Mada (UGM) in Indonesia. He earned a Bachelor of Science in Mechanical Engineering from UGM, a Master of Engineering in Manufacturing from the University of Malaya, and a Doctor of Engineering in Mechanical and Control Engineering from the Tokyo Institute of Technology in Japan. He currently teaches in the Department of Mechanical and Industrial Engineering at UGM in Yogyakarta, Indonesia. His research focuses on additive manufacturing systems. He can be contacted via email at herianto@ugm.ac.id.