ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Evaluating Trust in CNN Transfer Learning with Flower Image Classification via Heatmap-Based XAI

Rawesak Tanawongsuwan[1], Sukanya Phongsuphap[2] and Pattanasak Mongkolwat[3]

ABSTRACT

Convolutional neural networks (CNNs) have demonstrated impressive performance in image classification tasks but are often criticized for their black-box nature, which complicates understanding their decision-making and reliability. Transfer learning with pre-trained CNNs is a widely used approach for tasks with limited data. This study evaluates the performance and explainability of popular CNN models on flower image classification using two custom datasets, Flower-8-One and Flower-8-Zoom. Employing Explainable AI (XAI) techniques, such as Grad-CAM, this research visualizes CNN decision-making to uncover its alignment with human perception. A human study assesses trustworthiness by analyzing participants' confidence scores based on model visualizations. Results indicate strong CNN performance but highlight disparities between model-extracted features and human expectations. Among the models evaluated, Xception and Inception-v3 consistently earn higher trust ratings. These findings emphasize the necessity of XAI-driven evaluations to enhance trust and reliability in CNN-integrated systems, particularly in applications requiring human-computer interaction.

Article information:

Keywords: Convolutional Neural Networks, Explainable Artificial Intelligence, Transfer Learning, Trustworthiness, Flower Datasets, Image Classification

Article history:
Received: January 2, 2025
Revised: April 4, 2025
Accepted: July 7, 2025
Published: July 19, 2025 (Online)

DOI: 10.37936/ecti-cit.2025193.260320

1. INTRODUCTION

The success of AlexNet [1] in the ImageNet Large Scale Visual Recognition Challenge [2] demonstrated that deep learning networks can be developed and deployed to address significant computer vision problems, particularly in image classification [3]. Additionally, the use of transfer learning techniques with pre-trained CNN models has become very popular. This approach allows developers to quickly build a classifier and apply it to small datasets. Numerous studies have applied CNN-based transfer learning to various image classification tasks, including breast cancer detection [4, 5], glomerulus classification in nephropathology [6, 7], brain tumor detection [8], eye disease classification [9], paddy leaf disease identification [10], road surface crack detection [11], and COVID-19 detection [12].

The inherent complexity of CNNs often leads to their perception as black boxes. When selecting a CNN model for an image classification task, the process can feel like choosing from a series of opaque systems, with the added challenge of determining which one is most likely to deliver optimal results. Beyond performance, it is equally important to evaluate which models are reliable and trustworthy for specific needs.

This research was also motivated by our own experiences with CNNs. During the development of CNN classifiers using several models and applying them to our datasets, we found that transfer learning frequently resulted in strong performance for image classification tasks. However, when we analyzed how these models processed and classified input images, the results were surprising. The models' interpretations of the images often deviated significantly from our expectations, raising critical questions and diminishing our confidence in their reliability.

These challenges highlight the relevance of explainable AI (XAI). As described in [13], XAI methods aim to create models that are not only high-performing but also interpretable, enabling users to understand,

[1,2,3]The authors are with the Faculty of Information and Communication Technology, Mahidol University, Thailand, Email: rawesak.tan@mahidol.ac.th, sukanya.pho@mahidol.ac.th and pattanasak.mon@mahidol.ac.th

[1]Corresponding author: rawesak.tan@mahidol.ac.th

appropriately trust, and effectively manage AI systems. A survey on XAI by [14] identifies trustworthiness as a key goal, defining it as the confidence that a model will behave as intended when faced with a given problem. However, the survey also emphasizes that not all trustworthy models are inherently explainable, and trustworthiness is not easy to quantify. Fortunately, CNNs are better suited for explainability compared to many other machine learning models, as humans naturally excel at interpreting visual data.

This study asserts that although CNNs trained via transfer learning often achieve high classification accuracy, their internal decision-making processes do not always align with human intuition. Such misalignment can undermine user confidence in model predictions, especially in applications where interpretability and trust are critical. By combining visual explanation techniques with human evaluation, we propose that trustworthiness can be meaningfully assessed and differentiated across CNN architectures.

This study aims to evaluate and compare the performance of widely used CNN models trained through transfer learning for image classification tasks and to assess the trustworthiness of these models using visual explanation techniques, incorporating insights from human evaluation. The main contributions of this work are as follows:

1) To evaluate the effectiveness and trustworthiness of CNN models, we chose flowers as the primary classification objects. Initially, we reviewed several publicly available flower datasets [15-18], but these often contained images with inconsistent characteristics, such as single versus multiple flowers, partial versus complete flowers, and varying zoom levels. To ensure a systematic study and draw meaningful conclusions, we did not use these public datasets. Instead, we created two new custom flower datasets: Flower-8-One and Flower-8-Zoom. We carefully curated these datasets with well-defined characteristics to support evaluating model accuracy and trustworthiness through heatmap-based explanations.
2) We applied transfer learning to 19 pretrained CNN models, training them on the custom flower datasets. We tested these models on the same test dataset to systematically compare their classification accuracy and analyze their relative performance.
3) To the best of our knowledge, this is the first study to comprehensively examine the trustworthiness of a large number of CNN models, with a particular focus on 19 architectures. Our human evaluation study gathered insights into trustworthiness from a user perspective, providing valuable guidance on evaluating the reliability and trustworthiness of CNN models.
4) Although techniques like transfer learning, Grad-CAM, and participant-based scoring are well established, our integration of these components into a trust assessment framework for CNN-based flower classification offers a novel and practical application. The use of controlled, custom-built datasets and systematic human evaluation based on intuitive visual criteria (localization and consistency) provides a unique lens into how model behavior aligns with human expectations. This human-centered, explainability-driven methodology contributes new insights into trustworthiness assessment and offers actionable guidance for selecting CNN models in real-world, user-facing applications.

The findings of this research provide actionable insights into identifying CNN models that not only achieve high classification accuracy but also inspire confidence in their ability to perform as intended in classification tasks. This study offers practical guidance for selecting CNN models that balance technical performance with human-centered trustworthiness.

2. RELATED WORK

CNNs have seen remarkable advancements over the past decade, driving significant progress in various computer vision tasks. Researchers have developed numerous CNN architectures, each aiming to improve performance, efficiency, and interpretability. This section covers key concepts of CNNs, transfer learning, and the black box nature of these models.

2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs), a type of artificial neural network (ANN), are widely used in image analysis. Popular models like AlexNet [1], VGG [19], GoogLeNet [20], NASNet [21], DarkNet [22], EfficientNet [23], ShuffleNet [24], SqueezeNet [25], DenseNet [26], MobileNet [27], Inception [28], Xception [29], ResNet [30, 31], and Inception-ResNet [31] have achieved remarkable success in computer vision tasks. CNNs learn features from images through layers such as convolutional, pooling, and fully connected layers. Key design factors include the number of layers, filter sizes (smaller for local features, larger for global), and the learning process, which involves forward and backward propagation to optimize classification accuracy. Training CNNs requires large, diverse datasets and significant time and resources, producing pretrained models for reuse in various tasks.

2.2 Transfer Learning

Machine learning models are often task-specific and require retraining with new data for different tasks, limiting real-world applicability. Transfer learning solves this by leveraging knowledge from a previously learned task to improve performance on a

new one [32, 33]. In deep learning, pretrained CNN models trained on large datasets serve as effective starting points for smaller datasets, reducing data and computational demands [34-38]. This method has become central to CNN research, especially in image classification, with recent reviews exploring advancements in deep transfer learning [39, 40].

2.3 The Black Box Nature of CNNs and XAI

CNNs are often called "black boxes" due to their complex structures, making it hard to understand the link between input and output. This opacity underscores the need for XAI, which provides accessible explanations to clarify model workings and enhance trust. While some models are inherently interpretable, CNNs typically rely on post-hoc techniques, such as visual explanations, to highlight image regions influencing predictions and improve system transparency.

2.3.1 Visual Explanations of CNNs

Efforts to address the black box nature of CNNs often focus on visualization techniques that reveal what the network "sees" when making predictions. These methods include visualizing feature maps, attention mechanisms, or layer activations to help users understand how the network processes information. For instance, Activation Maximization [39] visualizes the features learned by hidden layers by generating images that maximize specific neuron activations, though it doesn't capture the complex interactions among neurons. Similarly, methods like DeconvNet [38] and Guided Backpropagation [40] offer different ways to visualize CNN behavior: DeconvNet generates feature-based visualizations, while Guided Backpropagation highlights regions in input images that influence network decisions, often creating more focused heatmaps.

A more advanced technique, Gradient-weighted Class Activation Mapping (Grad-CAM) [41], builds on Class Activation Maps (CAMs) [42] to generate heatmaps that highlight critical regions in an image for specific predictions. Grad-CAM achieves this by using the gradients of the predicted class score relative to the feature maps of the final convolutional layer. These gradients determine how much each feature map contributes to the output, and their weighted sum creates the heatmap. Grad-CAM provides high-resolution visualizations that enhance interpretability and localization, working effectively across different CNN architectures. An extension, Gradual Extrapolated Grad-CAM [43], improves heatmap sharpness by gradually applying gradient weights from the predicted class to other classes, though it increases computational overhead.

2.3.2 Trust Assessment of CNNs

Trust in CNNs depends on their interpretability and the ability to validate predictions. A study in [44] highlights that interpretability, a multifaceted concept, is crucial for trust as users struggle to rely on models they cannot understand.

Local Interpretable Model-agnostic Explanations (LIME) [45] enhances trust by treating models as black boxes and using local approximations to provide human-readable explanations for individual predictions. Similarly, heatmap-based XAI methods, like Saliency and Deconvolution, improve prediction confidence and reliability. A novel approach, Generative Augmentative Explanation, boosts confidence by tailoring heatmaps to specific datasets, underscoring the importance of customized XAI techniques [46].

The PRISM framework furthers trust by visualizing CNN decision-making through principal feature analysis, offering global insights into feature use, bias detection, and prediction validation [47]. Inspired by Grad-CAM's human evaluation study, which showed visual explanations enhance transparency, this work builds on heatmaps to reveal key image regions influencing decisions, fostering user confidence [41].

3. METHODOLOGY

This section outlines the creation of two custom flower datasets, the development of flower classifiers using transfer learning with pre-trained CNNs, and the use of visualization techniques to understand the decision-making process of CNNs and assess their reliability. Through this methodology, we aim to provide insights into the inner workings of CNNs, highlighting both their strengths and limitations in practical applications. By understanding these aspects, we hope to foster a deeper discussion about the reliability and trustworthiness of CNN models.

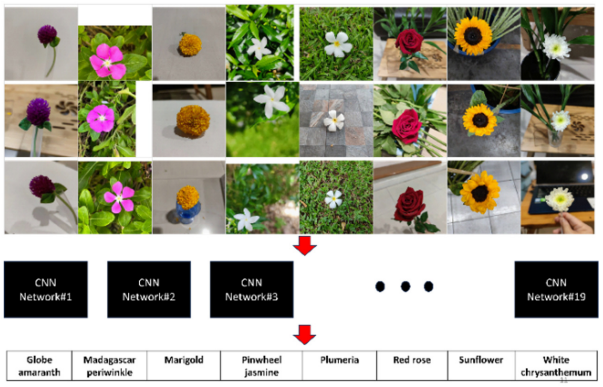

Fig.1: *Transfer learning of 19 pretrained CNNs with the Flower-8-One dataset.*

Fig.2: *Transfer learning of 19 pretrained CNNs with the Flower-8-Zoom dataset.*

3.1 Flower-8 Datasets

We created two flower datasets, each containing 1,120 images across eight flower types: Globe Amaranth, Madagascar Periwinkle, Marigold, Pinwheel Jasmine, Plumeria, Red Rose, Sunflower, and White Chrysanthemum (140 images per type).

- **Flower-8-One Dataset:** Features single flowers, often off-center, with surrounding context, designed to test CNNs' ability to localize and classify flowers.
- **Flower-8-Zoom Dataset:** Contains zoomed-in, centered flowers without surroundings, focusing on how CNNs analyze features and structures.

We divided each dataset into training, validation, and test subsets, labeled as Flower-8-One-Train/Test/Validate and Flower-8-Zoom-Train/Test/Validate. Table 1 summarizes the subset sizes, while Fig. **1** and Fig. **2** display sample images. The Flower-8-One and Flower-8-Zoom datasets used in this study are available from the corresponding author upon reasonable request, subject to the journal's data sharing policy.

Table 1: *Details of the Flower-8-One and Flower-8-Zoom datasets, each containing 1,120 images.*

Flower	Train	Validate	Test	Total
Globe Amaranth	100	20	20	140
Madagascar Periwinkle	100	20	20	140
Marigold	100	20	20	140
Pinwheel Jasmine	100	20	20	140
Plumeria	100	20	20	140
Red Rose	100	20	20	140
Sunflower	100	20	20	140
White Chrysanthemum	100	20	20	140
Total	**800**	**160**	**160**	**1120**

3.2 Transfer Learning Using Pretrained Networks

Pretrained networks, designed to classify images into broad categories, can be adapted to classify the specific eight flower types in our datasets. Instead of training models from scratch, we fine-tune pretrained networks, leveraging their learned features while customizing them for our flower datasets. This fine-tuning process involves modifying the final layers—replacing the fully connected and classification layers—to suit our task.

We train two sets of CNNs: one using the Flower-8-One-Train dataset and another using the Flower-8-Zoom-Train dataset. We use the Flower-8-One-Validate and Flower-8-Zoom-Validate datasets for hyperparameter tuning and performance evaluation.

3.2.1 Choosing Pretrained Networks

Pretrained networks, typically trained on large datasets like ImageNet, provide a strong starting point for transfer learning due to their generalized feature extraction capabilities. Our study utilizes 19 pretrained networks available in MATLAB, all trained on ImageNet. Table 2 provides details on these networks, including the number of parameters and input image sizes.

3.2.2 Replacing Final Layers

The final layers of pretrained networks, which generate class probabilities and predictions, are replaced with new layers tailored to classify our flower categories. This customization enables the networks to adapt their feature extraction capabilities to our datasets.

Table 2: *Details of 19 networks obtained from the MATLAB website [https://www.mathworks.com].*

♯	Name	Parameters (Millions)	Input Size
1	SqueezeNet	1.24	227-by-227
2	GoogLeNet	7	224-by-224
3	Inception-v3	23.9	299-by-299
4	DenseNet-201	20	224-by-224
5	MobileNet-v2	3.5	224-by-224
6	ResNet-18	11.7	224-by-224
7	ResNet-50	25.6	224-by-224
8	ResNet-101	44.6	224-by-224
9	Xception	22.9	299-by-299
10	InceptionResNet-v2	55.9	299-by-299
11	ShuffleNet	1.4	224-by-224
12	NASNet-Mobile	5.3	224-by-224
13	NASNet-Large	88.9	331-by-331
14	DarkNet-19	20.8	256-by-256
15	DarkNet-53	41.6	256-by-256
16	EfficientNet-b0	5.3	224-by-224
17	AlexNet	61	227-by-227
18	VGG-16	138	224-by-224
19	VGG-19	144	224-by-224

3.2.3 Data Augmentation

To enhance training and prevent overfitting, we resize input images to match network requirements and apply data augmentation. Specifically, we apply random vertical flips, translations up to 30 pixels, and

scaling up to 10%. These transformations increase dataset diversity and improve model robustness.

3.2.4 Training Pretrained Networks

We trained each network for 30 epochs using the SGDM optimizer with a learning rate of 0.0003 and a batch size of 10. We trained two sets of 19 CNNs: one with the Flower-8-One dataset and another on the Flower-8-Zoom dataset, all on NVIDIA DGX-A100 servers.

Despite using consistent hyperparameters, the networks achieved high training and validation accuracy with minimal loss, indicating effective learning without overfitting. Table 3 summarizes the validation accuracy (%) for these CNN models. Fig. 1 and Fig. 2 illustrate the 38 customized CNNs trained on our datasets.

Table 3: *Validation accuracy of 38 CNNs trained with Flower-8-One-Train and Flower-8-Zoom-Train sets.*

♯	Name	Flower-8-One Validation Accuracy (%)	Flower-8-Zoom Validation Accuracy (%)
1	SqueezeNet	100	98.12
2	GoogLeNet	100	100
3	Inception-v3	100	100
4	DenseNet-201	100	100
5	MobileNet-v2	100	100
6	ResNet-18	100	100
7	ResNet-50	100	100
8	ResNet-101	100	100
9	Xception	100	100
10	InceptionResNet-v2	88.75	83.12
11	ShuffleNet	100	100
12	NASNet-Mobile	98.12	100
13	NASNet-Large	99.38	100
14	DarkNet-19	100	100
15	DarkNet-53	100	100
16	EfficientNet-b0	100	100
17	AlexNet	100	100
18	VGG-16	100	100
19	VGG-19	100	100

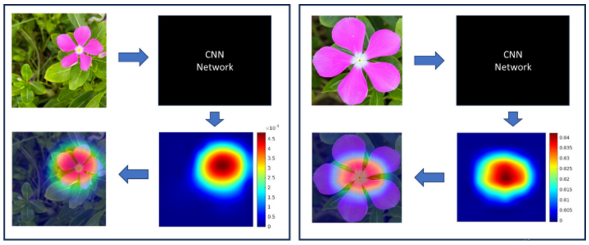

Fig.3: *The Grad-CAM heatmaps, overlaid on the original Madagascar Periwinkle flower images.*

3.3 Visualizing Network Predictions

Visualizing CNN predictions provides valuable insights into how networks make decisions, revealing their strengths and weaknesses. Grad-CAM, a widely used visualization method, balances positive gradient emphasis with some negative influence through weighting, offering robust interpretations and accurate localization of critical regions. This capability is especially useful for the Flower-8-One dataset, which focuses on flower localization and classification, and the Flower-8-Zoom dataset, which examines how CNNs analyze flower features. Grad-CAM heatmaps overlay directly on images, enabling compact and intuitive visualization of network predictions.

For our trust evaluation, participants can efficiently compare prediction results across test images using Grad-CAM's heatmap overlays. Its versatility also allows application across diverse CNN architectures without modification. For instance, Fig. 3 (Left) illustrates a Madagascar Periwinkle image from the Flower-8-One-Test set, where Grad-CAM highlights critical regions used by the network for classification. Similarly, Fig. 3 (Right) shows how Grad-CAM analyzes another Madagascar Periwinkle image from the Flower-8-Zoom-Test set, revealing how the network processes flower features.

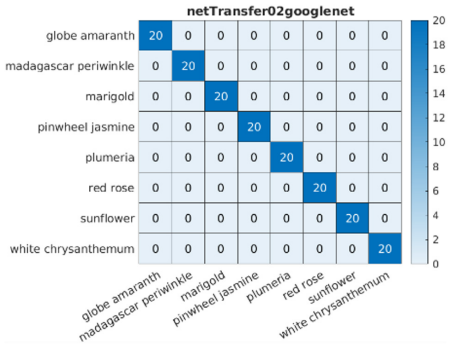

Fig.4: *A confusion matrix of network ♯2 (GoogLeNet) classifying images from the Flower-8-Zoom-Test set.*

4. EVALUATING CNN PERFORMANCE

We evaluated 19 networks trained on the Flower-8-One-Train set, validated with the Flower-8-One-Validate set, and tested on 160 images from the Flower-8-One-Test set. Similarly, we trained and validated another 19 networks on the Flower-8-Zoom-Train and Flower-8-Zoom-Validate sets, and tested them on 160 images from the Flower-8-Zoom-Test set.

We assessed the performance of all 38 networks using confusion matrices and accuracy rates. For example, Fig. 4 presents the confusion matrix for GoogLeNet (♯2) trained on Flower-8-Zoom-Train, showing 100% accuracy for this test set. Due to space constraints, Table 4 summarizes correct predictions and overall accuracy, representing the percentage of correctly classified test images.

5. EVALUATING TRUST OF CNNS

Despite their high classification accuracy, it is essential to assess the trustworthiness of CNN models.

Table 4: *Test accuracy of 38 CNNs trained with Flower-8-One-Train and Flower-8-Zoom-Train sets.*

♯	Name	Flower-8-One Test Accuracy (%)	Flower-8-Zoom Test Accuracy (%)
1	SqueezeNet	100	98.75
2	GoogLeNet	100	100
3	Inception-v3	100	100
4	DenseNet-201	100	100
5	MobileNet-v2	100	100
6	ResNet-18	100	100
7	ResNet-50	100	100
8	ResNet-101	100	100
9	Xception	100	100
10	InceptionResNet-v2	90	92.5
11	ShuffleNet	100	100
12	NASNet-Mobile	96.88	100
13	NASNet-Large	100	100
14	DarkNet-19	100	100
15	DarkNet-53	100	100
16	EfficientNet-b0	100	100
17	AlexNet	100	100
18	VGG-16	99.38	100
19	VGG-19	100	100

This section evaluates trustworthiness through visual explanation techniques and a human evaluation study. Grad-CAM visualizations generated heatmaps that highlight the areas CNNs focus on during classification. We presented these heatmaps to participants, who rated their confidence in the models' predictions.

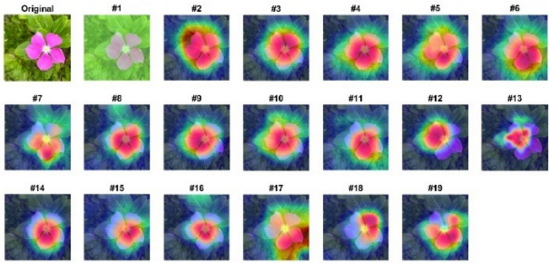

Fig.5: *A Madagascar Periwinkle (One) test image along with Grad-CAM heatmaps generated by 19 CNNs.*

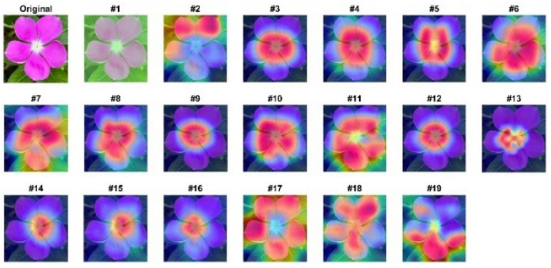

Fig.6: *A Madagascar Periwinkle (Zoom) test image along with Grad-CAM heatmaps generated by 19 CNNs.*

5.1 Human Study on CNN Trust Assessment

This study, approved by the Institutional Review Board (IRB) at the authors' institution, involved 30 participants aged 18–65 without color blindness. Participants provided informed consent and received a briefing on flower classification using CNNs, the assessment scope involving two groups of 19 trained models, heatmap visualizations, and confidence scoring criteria. Because the task centered on evaluating how well highlighted regions in the heatmaps aligned with recognizable flower objects, it did not require technical knowledge of CNNs or AI. We provided all participants with clear, standardized instructions and visual examples explaining how to score each heatmap on the defined set of criteria.

We conducted the evaluation in two stages:

1. **Flower-8-One-Test study:** Participants reviewed 160 image sets, comprising 20 original test images for each of the eight flower types (8 sets) and their corresponding Grad-CAM heatmaps generated by 19 CNNs (152 sets). Participants used the interface shown in Fig. 7 to assign confidence scores based on the visual explanations.
2. **Flower-8-Zoom-Test study:** The second stage followed the same procedure as the first but used the Flower-8-Zoom-Test dataset. Participants again reviewed 160 image sets, focusing on the networks' ability to analyze structural flower features.

5.2 Trust Assessment from Visualization of CNN Predictions

We selected Grad-CAM to visualize CNN predictions for both datasets. It generates heatmaps highlighting critical areas influencing model decisions. Fig. 5 and Fig. 6 present examples of Grad-CAM visualizations for a Madagascar Periwinkle image from the Flower-8-One-Test and Flower-8-Zoom-Test sets, respectively. These heatmaps help assess the networks' localization and feature analysis capabilities, showing variations in focus areas across different CNNs.

Participants evaluated the networks on two criteria:

- **Localization accuracy:** Assessing how well heatmaps highlight the flower's position and shape.
- **Consistency:** Evaluating uniformity in heatmap results across the 20 test images for each flower type.

Each participant scored the CNNs using a five-point scale: 1 (very low), 2 (low), 3 (medium), 4 (high), and 5 (very high). The analysis compared confidence scores across networks. For example, the following observations were noted:

- For Globe Amaranth (One), EfficientNet-b0 achieved higher scores than InceptionResNet-v2 due to better localization and consistent heatmaps (Fig. 8 a1-a3).

- For Plumeria (Zoom), DenseNet-201 outperformed VGG-16, offering more precise and more consistent focus on flower structures (Fig. 9 a1-a3).

The findings highlight how explainability techniques, such as Grad-CAM, can provide transparency into CNN decision-making, fostering trust by enabling users to verify the basis for its predictions.

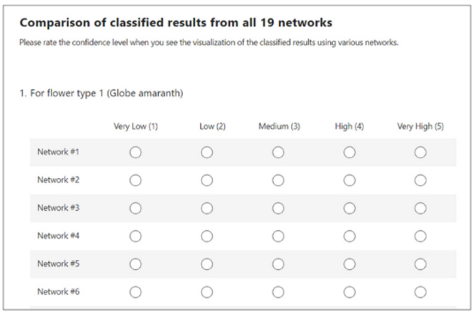

Fig.7: *The scoring interface for participants to enter their confidence scores.*

5.3 Degree of Trust in CNNs

This section presents the analysis of participant confidence scores to evaluate trust in CNN models for flower classification tasks.

5.3.1 Trust Analysis from Flower-8-One-Test Set

In Table 5, we summarize the mean confidence scores for 19 CNN models that classified test flower images from the Flower-8-One-Test set and use boxes to highlight the top five networks with the highest mean scores for each flower type. For example:

- **Madagascar Periwinkle:** DarkNet-19 (4.33), Inception-v3 (4.07), ResNet-101 (3.93), Xception (3.90), and DenseNet-201 (3.87).
- **Plumeria:** VGG-16 (4.67), Xception (4.33), ShuffleNet (4.33), ResNet-101 (4.30), and EfficientNet-b0 (4.30).

Fig. 8 shows test images of Globe Amaranth, Madagascar Periwinkle, Marigold, and Pinwheel Jasmine (One version), along with the heatmaps of the networks with the highest and the 2nd lowest mean confidence scores.

Networks with higher confidence scores demonstrated effectiveness in localizing flower objects and consistency across test images of the same flower type.

It is important to note that the networks with the lowest scores are not displayed in Fig. 8. In most instances, SqueezeNet received the lowest mean scores, as its heatmaps typically displayed minimal focus areas, characterized by sparse red pixels. To provide a broader range of examples, the analysis includes networks with the next-lowest scores instead.

When considering overall performance, the top five most trusted CNNs based on the mean scores across all flower types were Xception (4.27), Inception-v3 (4.16), ResNet-101 (4.12), EfficientNet-b0 (4.11), and ShuffleNet (4.03).

5.3.2 Trust Analysis from Flower-8-Zoom-Test Set

Table 6 shows the mean confidence scores for the same 19 CNNs, this time classifying images from the Flower-8-Zoom-Test set. For instance:

- **Madagascar Periwinkle:** MobileNet-v2 (4.50), DenseNet-201 (4.27), Xception (4.13), Inception-ResNet-v2 (4.13), and NASNet-Mobile (4.07).
- **Plumeria:** DenseNet-201 (4.43), Inception-ResNet-v2 (4.37), NASNet-Mobile (4.37), Inception-v3 (4.33), and Xception (4.33).

These scores reflect the networks' ability to analyze and capture flower features consistently.

Similarly, Fig. 9 omits the networks with the lowest scores. SqueezeNet consistently received the lowest mean scores in all cases, so the analysis includes networks with the next-lowest scores instead.

The overall top five most trusted CNNs for the Flower-8-Zoom-Test set were NASNet-Mobile (4.23), Xception (4.13), DenseNet-201 (4.09), Inception-v3 (3.99), and MobileNet-v2 (3.89).

6. DISCUSSION

This study examined the performance and trustworthiness of 19 CNN models trained through transfer learning for flower classification tasks. The Flower-8-One dataset enabled the evaluation of CNNs' ability to localize and classify flowers, while the Flower-8-Zoom dataset provided insights into how CNN models analyze flower features.

The study provided insights into CNN performance and trustworthiness:

- **Performance:** The results demonstrated that CNNs achieved near-perfect accuracy rates for flower classification, except for Inception-ResNet-v2, which showed relatively lower performance. These findings confirm the effectiveness of CNNs as robust models for image classification tasks.
- **Visualization for Trust:** Visualization techniques, such as Grad-CAM, played a crucial role in helping users understand CNN decision-making, effectively mitigating the black-box nature of these models.
- **Human Alignment:** While CNNs deliver high accuracy, their analysis methods may not align with human expectations, influencing trust.
- **Top Performers in Flower-8-One:** For the Flower-8-One dataset, models like Xception and Inception-v3 consistently focused on flower regions, aiding in accurate classification. Examination of test images and corresponding heatmaps, as shown in Fig. 8, revealed that the top-performing networks maintained a high degree of accuracy in localizing flowers and delivering consistent results. The most trusted models for this

dataset were Xception, Inception-v3, ResNet-101, EfficientNet-b0, and ShuffleNet.

- **Top Performers in Flower-8-Zoom:** In the Flower-8-Zoom dataset, networks like DenseNet-201 and NASNet-Mobile excelled at analyzing flower structures and demonstrated consistency across test data. Visualized heatmaps, as illustrated in Fig. 9, highlighted the networks' ability to focus on critical flower features effectively. The most trusted models for this dataset were NASNet-Mobile, Xception, DenseNet-201, Inception-v3, and MobileNet-v2.
- **Consistent Top Performers:** Xception and Inception-v3 emerged as consistently top-performing models across both datasets. These findings underscore their reliability and effectiveness in localizing and analyzing flower structures, reinforcing their suitability for flower classification tasks. This study highlights the value of visualization tools and user-centered evaluations in assessing the trustworthiness of CNNs

7. CONCLUSION

This study systematically evaluated the performance and trustworthiness of 19 convolutional neural network (CNN) models trained using transfer learning on two custom flower datasets, Flower-8-One and Flower-8-Zoom. By employing visual explanation techniques such as Grad-CAM, we provided insights into the decision-making processes of these models, bridging the gap between technical accuracy and human-centered trust. The study revealed that while most CNN models achieved high classification accuracy, their trustworthiness varied significantly based on localization accuracy and consistency as perceived by human evaluators. Models such as Xception and Inception-v3 emerged as not only high-performing but also consistently trustworthy across various flower types. These findings highlight the importance of explainability techniques in fostering confidence in CNN-based classification systems and offer actionable guidance for model selection in applications requiring both accuracy and reliability.

Future research can extend this study by exploring the reliability of trusted networks across diverse object types and domains, such as medical imaging or agriculture. Expanding dataset diversity and incorporating real-world complexities like occlusions or mixed arrangements will improve model robustness. Investigating advanced explainable AI techniques and analyzing network architectures could provide deeper insights into CNN decision-making and effectiveness. Additionally, developing interactive trust metrics for dynamic user engagement with predictions and heatmaps will enhance interpretability and foster trust in practical applications.

ACKNOWLEDGEMENT

In this work, training CNN models on DGX-A100 servers was supported by Mahidol University and the Office of the Ministry of Higher Education, Science, Research, and Innovation under the Reinventing University project: the Center of Excellence in AI-Based Medical Diagnosis (AI-MD) sub-project.

AUTHOR CONTRIBUTIONS

Conceptualization, R.T. and S.P.; methodology, R.T.; software, R.T.; validation, R.T., S.P., and P.M.; formal analysis, R.T. and S.P.; investigation, R.T. and S.P.; resources, R.T., S.P., and P.M.; data curation, R.T.; writing—original draft preparation, R.T.; writing—review and editing, S.P. and P.M.; visualization, R.T. and P.M.; supervision, R.T.; project administration, R.T. All authors have read and agreed to the published version of the manuscript.

References

[1] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015.

[3] N. K. Sarkar, M. M. Singh, A. Pradesh and U. Nandi, "Recent Researches on Image Classification Using Deep Learning Approach," *International Journal of Computing and Digital Systems*, vol. 12, pp. 1357-1374, 2022.

[4] C. R. Prasad, B. Arun, S. Amulya, P. Abboju, S. Kollem and S. Yalabaka, "Breast Cancer Classification using CNN with Transfer Learning Models," *International Conference for Advancement in Technology (ICONAT)*, pp. 1-5, 2023.

[5] L. Abdelrahman, M. Al Ghamdi, F. Collado-Mesa and M. Abdel-Mottaleb, "Convolutional neural networks for breast cancer detection in mammography: A survey," *Computers in Biology and Medicine*, vol. 131, no. 104248, 2021.

[6] J. Gallego, A. Pedraza, S. Lopez, G. Steiner, L. Gonzalez, A. Laurinavicius and G. Bueno, "Glomerulus Classification and Detection Based on Convolutional Neural Networks," *Journal of Imaging*, vol. 4, no. 1:20, 2018.

[7] P. Chagas, L. Souza, I. Araújo, N. Aldeman, A. Duarte, M. Angelo, W. L. C. dos-Santos and L. Oliveira, "Classification of glomerular hypercellularity using convolutional features and support vector machine," *Artificial Intelligence in Medicine*, vol. 103, no. 101808, 2020.

[8] K. N. Rao, O. I. Khalaf, V. Krishnasree, A. S. Kumar, D. M. Alsekait, S. S. Priyanka, A. S. Alattas and D. S. AbdElminaam, "An efficient brain tumor detection and classification using pre-trained convolutional neural network models," *Heliyon*, vol. 10, no. 17, pp. e36773, 2024.

[9] D. Helen, and S. Gokila, "EYENET: An Eye Disease Detection System using Convolutional Neural Network," *International Conference on Edge Computing and Applications (ICECAA)*, pp. 839-842, 2023.

[10] A. Islam, N. R. Shuvo, M. Shamsojjaman, S. Hasan, S. Hossain and T. Khatun, "An Automated Convolutional Neural Network Based Approach for Paddy Leaf Disease Detection," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021.

[11] N. H. T. Nguyen, S. Perry, D. Bone, H. T. Le and T. T. Nguyen, "Two-stage convolutional neural network for road crack detection and segmentation," *Expert Systems with Applications*, vol. 186, no. 115718, 2021.

[12] A. Narin, C. Kaya and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," *Pattern Analysis and Applications*, vol. 24, no. 3, pp. 1207-1220, 2021.

[13] D. Gunning, "Explainable artificial intelligence (xAI)," *Defense Advanced Research Projects Agency (DARPA)*, 2017.

[14] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82-115, 2020.

[15] K. Tung, "Flowers Dataset," Harvard Dataverse, 2020.

[16] "Flowers Recognition," Source. [Online]. Available: `https://www.kaggle.com/datasets/alxmamaev/flowers-recognition`.

[17] "Flower Datasets," Source. [Online]. Available: `https://www.robots.ox.ac.uk/~vgg/data/flowers/`.

[18] "Flowers dataset," Source. [Online]. Available: `https://www.kaggle.com/datasets/imsparsh/flowers-dataset`.

[19] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 3rd International Conference on Learning Representations (ICLR), pp. 1-14, 2015.

[20] C. Szegedy, L. Wei, J. Yangqing, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 1-9, 2015.

[21] B. Zoph, V. Vasudevan, J. Shlens and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8697-8710, 2018.

[22] J. Redmon. "Darknet: Open Source Neural Networks in C," https://pjreddie.com/darknet.

[23] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105-6114, 2019.

[24] X. Zhang, X. Zhou, M. Lin and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848-6856, 2018.

[25] N. I. Forrest, W. M. Matthew, A. Khalid, H. Song, J. D. William and K. Kurt, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size," *CoRR*, 2016.

[26] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 2261-2269, 2017.

[27] 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, pp. 4510-4520, 2018.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 2818-2826, 2016.

[29] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 1800-1807, 2017.

[30] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778, 2016.

[31] C. Szegedy, S. Ioffe, V. Vanhoucke and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4278–4284, 2017.

[32] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," in *IEEE Transactions on Knowledge*

and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[33] L. Torrey and J. Shavlik, "Transfer learning," *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242-264, 2010.

[34] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng and T. Darrell, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," *31st International Conference on Machine Learning*, pp. 647-655, 2014.

[35] M. Oquab, L. Bottou, I. Laptev and J. Sivic, "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 1717-1724, 2014.

[36] A. S. Razavian, H. Azizpour, J. Sullivan and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Columbus, OH, USA, pp. 512-519, 2014.

[37] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *2nd International Conference on Learning Representations (ICLR)*, 2014.

[38] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," *European Conference on Computer Vision (ECCV)*, pp. 818-833, 2014.

[39] D. Erhan, Y. Bengio, A. Courville and P. Vincent, "Visualizing Higher-Layer Features of a Deep Network," Technical Report, Univeristé de Montréal, 2009.

[40] J. Springenberg, A. Dosovitskiy, T. Brox and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," *International Conference on Learning Representations (Workshop)*, 2015.

[41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 618-626, 2017.

[42] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning Deep Features for Discriminative Localization," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 2921-2929, 2016.

[43] T. Szandała, "Enhancing Deep Neural Network Saliency Visualizations With Gradual Extrapolation," *IEEE Access*, vol. 9, pp. 95155-95161, 2021.

[44] Z. C. Lipton, "The Mythos of Model Interpretability," *ACM Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[45] M. T. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," *International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[46] E. Tjoa, H. J. Khok, T. Chouhan and C. Guan, "Enhancing the confidence of deep learning classifiers via interpretable saliency maps," *Neurocomputing*, vol. 562, no. 126825, 2023.

[47] T. Szandała, "Unlocking the black box of CNNs: Visualising the decision-making process with PRISM," *Information Sciences*, vol. 642, no. 119162, 2023.

Table 5: *Mean confidence scores from 19 networks classifying the Flower-8-One-Test set. Boxes highlight the top 5 mean confidence scores for each flower test set and the top 5 most trusted models.*

#	Name	Globe Amaranth	Madagascar Periwinkle	Marigold	Pinwheel Jasmine	Plumeria	Red Rose	Sunflower	White Chrysanthemum	Overall
1	SqueezeNet	1.63	1.03	1.03	1.20	1.23	1.37	1.17	1.33	1.25
2	GoogLeNet	3.20	3.27	3.27	2.47	3.37	3.57	2.87	2.40	3.05
3	Inception-v3	4.10	**4.07**	**4.33**	**3.77**	4.17	**4.27**	**4.50**	4.10	**4.16**
4	DenseNet-201	3.57	**3.87**	3.90	3.10	4.00	**4.30**	**4.40**	3.87	3.88
5	MobileNet-v2	3.43	3.53	3.97	3.10	3.43	3.93	4.13	3.53	3.63
6	ResNet-18	3.63	3.83	4.07	3.13	3.83	3.73	3.87	3.73	3.73
7	ResNet-50	**4.27**	3.43	4.27	3.27	3.97	**4.37**	3.50	**4.33**	3.93
8	ResNet-101	4.17	**3.93**	**4.47**	3.30	**4.30**	**4.40**	3.87	**4.50**	**4.12**
9	Xception	**4.23**	**3.90**	**4.40**	**4.27**	**4.33**	**4.30**	**4.40**	**4.33**	**4.27**
10	InceptionResNet-v2	1.80	3.67	1.83	2.20	4.07	3.83	4.00	2.23	2.95
11	ShuffleNet	**4.27**	3.43	**4.33**	**3.43**	**4.33**	3.73	**4.60**	4.07	**4.03**
12	NASNet-Mobile	2.67	3.83	2.47	**3.73**	4.13	3.23	3.73	2.17	3.25
13	NASNet-Large	**4.23**	3.47	3.53	**3.63**	4.20	3.53	3.70	**4.33**	3.83
14	DarkNet-19	**4.47**	**4.33**	4.10	2.40	4.17	3.17	4.00	3.87	3.81
15	DarkNet-53	3.60	3.67	2.67	3.00	3.73	3.93	**4.33**	3.70	3.58
16	EfficientNet-b0	**4.63**	3.73	**4.50**	2.93	**4.30**	4.07	4.27	**4.47**	**4.11**
17	AlexNet	3.47	3.37	3.40	1.43	4.07	2.87	4.07	3.17	3.23
18	VGG-16	3.60	3.10	3.10	1.13	**4.67**	4.03	4.30	4.10	3.50
19	VGG-19	3.53	3.23	4.17	1.67	1.83	2.30	3.53	4.17	3.05

Table 6: *Mean confidence scores from 19 networks classifying the Flower-8-Zoom-Test set. Boxes highlight the top 5 mean confidence scores for each flower test set and the top 5 most trusted models.*

#	Name	Globe Amaranth	Madagascar Periwinkle	Marigold	Pinwheel Jasmine	Plumeria	Red Rose	Sunflower	White Chrysanthemum	Overall
1	SqueezeNet	1.30	1.03	1.13	1.20	1.03	1.13	1.03	1.20	1.13
2	GoogLeNet	2.87	1.90	2.70	3.50	3.10	3.47	3.20	3.03	2.97
3	Inception-v3	3.03	3.93	**3.87**	**4.47**	**4.33**	**3.90**	**4.10**	**4.30**	**3.99**
4	DenseNet-201	**3.20**	**4.27**	**3.53**	3.97	**4.43**	**4.27**	**4.60**	**4.43**	**4.09**
5	MobileNet-v2	**3.67**	**4.50**	**3.53**	**4.20**	3.47	3.70	**4.23**	3.83	**3.89**
6	ResNet-18	**3.60**	3.97	**3.63**	3.97	3.03	**3.87**	**4.67**	3.93	3.83
7	ResNet-50	3.00	2.73	3.13	4.17	3.40	3.33	3.77	3.73	3.41
8	ResNet-101	2.77	3.53	3.00	3.90	3.97	3.63	3.77	3.90	3.56
9	Xception	**3.63**	**4.13**	**4.27**	**4.40**	**4.33**	**4.03**	3.87	**4.33**	**4.13**
10	InceptionResNet-v2	1.53	**4.13**	2.13	**4.27**	**4.37**	3.27	**4.07**	2.20	3.25
11	ShuffleNet	2.60	2.90	2.80	3.07	2.73	3.17	2.83	2.57	2.83
12	NASNet-Mobile	**4.17**	**4.07**	**4.63**	**4.40**	**4.37**	**4.10**	3.87	**4.27**	**4.23**
13	NASNet-Large	2.27	3.33	2.57	3.40	2.87	2.37	2.33	2.37	2.69
14	DarkNet-19	2.03	3.67	2.87	3.80	4.10	2.80	3.47	**4.20**	3.37
15	DarkNet-53	2.33	3.07	2.23	3.10	4.07	3.00	2.50	3.33	2.95
16	EfficientNet-b0	2.87	3.40	2.90	3.90	3.73	3.67	3.30	4.00	3.47
17	AlexNet	1.77	2.83	2.03	1.40	3.00	1.63	3.10	1.77	2.19
18	VGG-16	2.50	3.40	3.50	3.23	1.97	3.10	2.27	1.90	2.73
19	VGG-19	2.50	1.90	3.00	3.57	2.63	2.17	3.60	2.03	2.68

Fig.8: *Test images of Globe Amaranth, Madagascar Periwinkle, Marigold, and Pinwheel Jasmine (One version), along with the heatmaps of the networks with the highest and the 2^{nd} lowest mean confidence scores.*

Fig.9: *Test images of Plumeria, Red Rose, Sunflower, and White Chrysanthemum (Zoom version), along with the heatmaps of the networks with the highest and the* 2^{nd} *lowest mean confidence scores.*

Rawesak Tanwongsuwan is a faculty member in the faculty of Information and Communication Technology, Mahidol University, Thailand. His main research interests include computer vision, image processing, and computer graphics.

Pattanasak Mongkolwat is a faculty member in the faculty of Information and Communication Technology, Mahidol University, Thailand. His main research interests include radiology information system, integrating the healthcare enterprise, digital imaging and communications in medicine, health level 7 (HL7), hospital information system, and picture archiving and communication systems.

Sukanya Phongsuphap is a faculty member in the faculty of Information and Communication Technology, Mahidol University, Thailand. Her main research interests include image processing, pattern recognition, digital signal processing, and heart rate variability.