IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 21 June 2023, accepted 11 July 2023, date of publication 19 July 2023, date of current version 11 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3296792

RESEARCH ARTICLE

Explainable Deep Learning Models With Gradient-Weighted Class Activation Mapping for Smart Agriculture

LUYL-DA QUACH[1], KHANG NGUYEN QUOC[1], ANH NGUYEN QUYNH[1], NGUYEN THAI-NGHE[2], AND TRI GIA NGUYEN[3], (Senior Member, IEEE)
[1]FPT University, Can Tho Campus, Can Tho 900000, Vietnam
[2]Faculty of Information Systems, College of ICT, Can Tho University, Can Tho 910000, Vietnam
[3]FPT University, Quy Nhon Campus, Quy Nhon 551170, Vietnam

Corresponding author: Nguyen Thai-Nghe (ntnghe@cit.ctu.edu.vn)

ABSTRACT Explainable Artificial Intelligence is a recent research direction that aims to explain the results of the Deep learning model. However, many recent research need to go into depth in evaluating the effectiveness of deep learning models in classifying image objects. For that reason, the research proposes two stages in the process of applying Explainable Artificial Intelligence, including: (1) assessing the accuracy of the deep learning model through evaluation methods, (2) using Grad-CAM for model interpretation aims to evaluate the feature detection ability of an image when recognized by deep learning models. The deep learning models included in the evaluation included VGG16, ResNet50, ResNet50V2, Xception, EfficientNetV2, InceptionV3, DenseNet201, MobileNetV2, MobileNet, NasNetMobile, RegNetX002, and InceptionResNetV2 on our updated VegNet dataset is available at: https://www.kaggle.com/datasets/enalis/tomatoes-dataset. The results show that the MobieNet model has high accuracy but less reliability than EfficientNetV2S and Xception. However, MobileNetV2's accuracy is the highest when considering the ratio match rate. The research results contribute to the construction of intelligent agricultural support systems (using automatic fruit-picking robots, removing poor-quality fruits,...) from the results of the Explainable AI model to be able to use the optimal deep learning model in processing.

INDEX TERMS Explainable artificial intelligence, XAI, agriculture, grad-CAM, deep learning, explainable AI.

I. INTRODUCTION

Explainable Artificial Intelligence (XAI) is a new branch of artificial intelligence. XAI sets of methods and tools that can make machine learning models easier to understand for humans. The purpose of XAI is to explain the results produced by constructing machine learning models [1]. However, the question arises about the reliability of Deep learning models (DL) for the features used for classification. Therefore, XAI is used to evaluate the accuracy of the DL model on the VegNet dataset [2] with the Gradient-weighted Class Activation Mapping (Grad-CAM) technique [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

In recent times, artificial intelligence (AI) has become increasingly popular and supports people a lot in life, especially in medicine, agriculture, and decision support,... In the medical field, research related to the medical image registration process [4], cancer diagnosis [5], disease detection [6], [7],... Meanwhile, the agricultural field has also had much research related to identification and detection, such as diseases in poultry [8], [9], diseases on crops [10], diseases in shrimp [11],... All this research shows that AI systems are essential and save more time in classification and diagnosis than humans. However, research [12], [13] shows that the critical issue of AI is decision support and data transparency in statistics, which makes for easy decision-making, and the first step towards remedy is an essential aspect of accountability. Although the DL model is strongly developed, the internal architecture is nested, making it difficult to explain or visualize the problem, often called a "black box". In the research [14], the authors suggested that deep learning models can be easily fooled by attacks in the physical world, such as environmental conditions, input data dependence, black and white labels,... This makes the development of XAI in the recent period [15]. XAI was born to meet the system's transparency, interpretability and explainability, known as the "white box" [16]. XAI shows a strong impact through much research using different techniques such as CAM [17], LPR [18], Grad-CAM [19], Guided backpropagation [20],... All show the need for XAI in decision support and transparency in data interpretation.

Research on XAI implementation in agriculture is little, but it shows the effectiveness of data processing in agriculture. Research [21] shows the importance of food security in the Sustainable Development Goals of the United Nations. Research [22] used XAI to explain tomato leaf diseases. Research [23] estimates nitrogen in wheat from proximal hyperspectral data using XAI and machine learning. Research shows the importance of XAI not only in medical and military research and agriculture and contributes to a certain development. In particular, according to statistics, few research uses XAI in agriculture.

In summary, through the survey, the contribution of the research is obtained through:

- Determine the accuracy of the DL model in classifying data on the VegNet dataset. The dataset used evaluates the data set regarding the severity of food safety in agriculture.
- Use Grad-CAM to visualize diagnostic data on the VegNet dataset with existing DL models.
- Re-evaluate the accuracy and propose the construction of XAI models for the agricultural sector in the future.

To achieve the above objectives, the study proposes a model consisting of 2 stages:

- Stage 1: Classifying the quality of tomatoes based on images using the DL model and evaluating the model's effectiveness according to the scales in image classification. During this process, the research team added additional datasets to the VegNet dataset to balance the amount of trained/test/validation data.
- Stage 2: Local explainability of models using Grad-CAM and compare the reliability of models with each other. Based on the resulting visualization, the research evaluates the accuracy of the prediction process of the DL model based on image segmentation.

II. RELATED WORK

DL models used in research can learn, recognize, classify and detect objects in images. They can vary in structure, number and type of deep neural network layers, model depth and complexity. The general property of models is that deeper and more complex models usually have higher accuracy on large data sets, but they require more time and resources to train and use. Specifically, the models used include VGG16 [24], ResNet50 [25], ResNet50V2 [26], Xception [27], EfficientNetV2 [28], InceptionV3 [29], DenseNet201 [30], MobileNetV2 [31], MobileNet [32], NasNetMobile [33], RegNetX002 [34], InceptionResNetV2 [35]. As mentioned above, the DL model provides accuracy, but the problem is that the processing consists of many layers, leading to a transparency problem that needs to be explained.

For that reason, XAI has made strides in building and developing tools to support interpreting the results of the DL model, which makes XAI techniques in data processing after the classification process more and more developed. XAI has created many successes in various fields, such as transportation systems that help reduce vehicle accidents with severe consequences [36]; The Healthcare system is based on clinically interpretable predictions to avoid the process of giving false support leading to deaths [37]; legal systems that reduce costs, risk of recidivism or automated decision-making [38]; the financial system automatically makes decisions that aid in interpretation, making it more audit-friendly, detecting illegal transactions [39], [40]; The military system starting from the DAPRA project demonstrates the development and reliability of XAI in supporting important decisions in the military [41]. In the research, it is found that the prediction is wrong in any field. Any sector will affect the cost [42]. Therefore, the XAI approach that needs to be taken in the agricultural sector, especially food-level classification, is fundamental.

With the development of XAI, there are many possibilities for developing research directions in applying XAI in the agricultural field. In this context, based on the results and success of XAI, the research team proposes to develop and apply the results of XAI in the agricultural sector, initially using Grad-CAM to assess the accuracy of DL models in the VegNet dataset.

III. METHODOLOGY

A. OVERALL METHODOLOGY

The research process was divided into two different phases. In phase 1, the research team began to collect data to add to the VegNet dataset on the tomato damage data set. Designated labels categorize the data, including Unripe, Ripe, Old and Damaged. After that, the data is normalized to a standard dataset form and divided the data set into 3 evaluators to serve the training/testing of DL and XAI models. Through the training process, based on fundamental metrics, we will perform tests on the trained model to evaluate the effectiveness of the DL models (FIGURE. 1). Stage 2, uses the Grad-CAM algorithm to explain the results of the DL model by identifying image features in detected regions and the importance of features. This is to evaluate the image recognition of DL or black box models after detecting features to make accurate predictions. In addition, to evaluate the completeness of the model based on Grad-CAM, the research evaluates all the image features in the test dataset on the models to clarify the differences in the model's learned features on each label (FIGURE. 2).

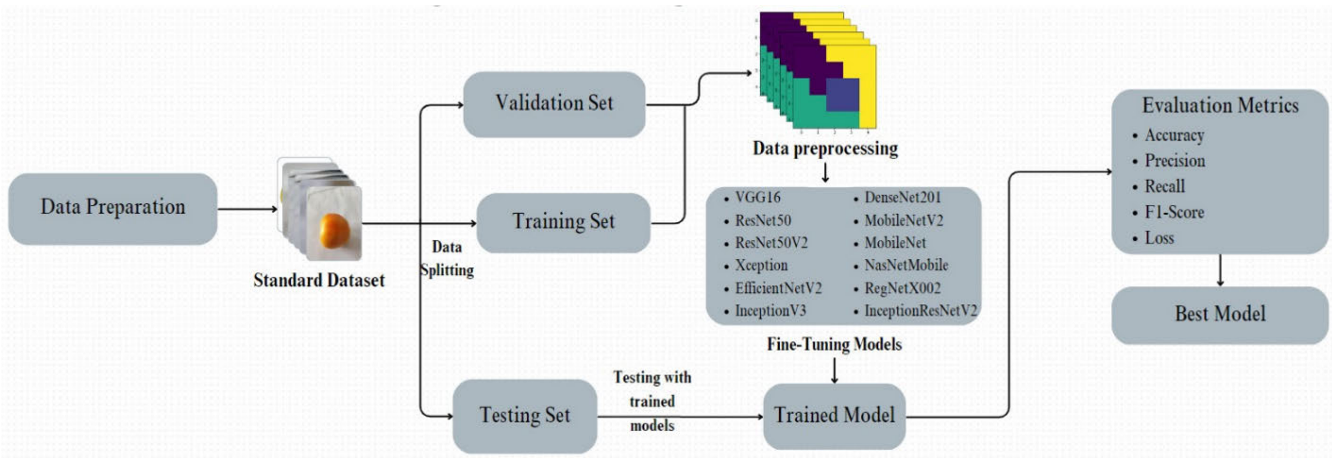


FIGURE 1. The process of identifying tomato state images by DL Models and evaluating the model with traditional metrics.

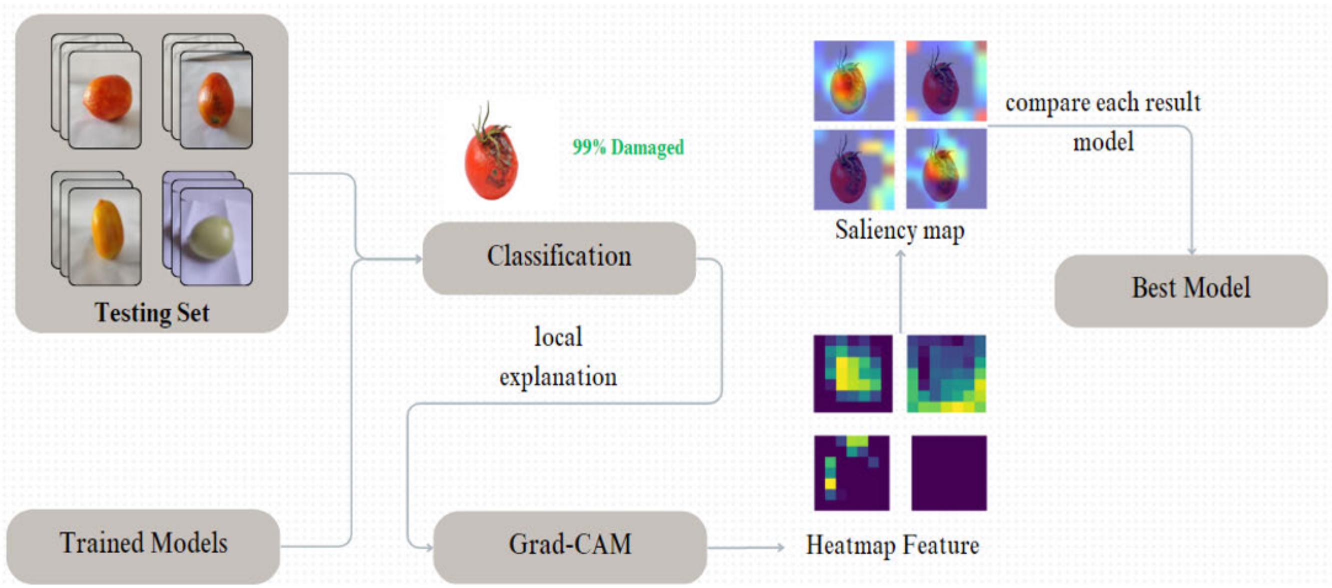


FIGURE 2. The process of interpreting DL model and evaluating the effectiveness of the model based on the Saliency Map based on Grad-CAM.

B. DATA PREPARATION

The image dataset about the state of the tomato in the VegNet [2] dataset has 4 states: Unripe, Ripe, Old and Damaged. However, this data set significantly differs between classes, so the research team collected more data sets. The results are collected with more than 2000 images from many different sources, such as 60% from the VegNet dataset, 30% from Kaggle and 10% from other sources. Due to the heterogeneous data size, the research resized it to 256 × 256 pixels (the standard size of the VegNet dataset) to create a standard dataset for future research. The research team creates the characteristics of the current dataset to create a white background. It divides the data set into 3 sets, including a training set, validation set and testing set in the ratio 6:2:2. The amount of data is detailed in TABLE 1.

TABLE 1. The statistics of the number of images in each class of the dataset.

Classes	Training Set	Validation Set	Testing Set
Unripe	296	94	93
Ripe	357	114	114
Old	317	103	103
Damaged	250	98	97
Total	**1220**	**409**	**407**

C. DATA PREPROCESSING

During the training in stage 1, the data is processed according to the following steps: (1) the data is resized to the standard size of the DL models, respectively 224 × 224 or 299 × 299 pixels; (2) For the training set, the data is applied augmentation to increase the image size such as random flip,

TABLE 2. Information about the DL models used by us.

DL model	Params	Memory Size	Publication Year
VGG16 (VG16)	15.2	62.3	2014
ResNet50 (RES1)	25.6	115	2015
ResNet50V2 (RES2)	25.6	114	2017
Xception (XCEP)	22.9	104	2017
EfficientNetV2 (EFN2)	21.6	93.9	2021
InceptionV3 (INV3)	23.9	108	2015
DenseNet201(D201)	20.2	94.3	2016
MobileNetV2 (MON2)	3.57	24.1	2018
MobileNet(MON1)	4.28	24.6	2017
NasNetMobile(NANM)	5.35	30.8	2018
RegNetX002(REN2)	2.71	13.7	2020
InceptionResNetV2(IRN2)	23.9	108	2016

TABLE 3. Invariant hyperparameter for training model.

Hyperparameters	Setting
Learning Rate	10^4
Algorithm Optimization	Adam
Epoch	100
Batch size	32

rotate and zooming (FIGURE. 3); (3) Image data is converted to vector and scaled again with the ratios [-1;1], [0,1] or keep [0;255]. After being processed, the data will be fed into the DL model to train and check the accuracy of the model. For Stage 2, the research uses the test data set to process and put the explanatory process by Grad-CAM technique into the evaluation model. The processing is illustrated in FIGURE. 4.

D. FINE-TUNING DEEP LEARNING MODELS AND TRAINING PROGRESS

The DL model in computer vision is evaluated more effectively than the traditional machine learning model [43], [44]. Therefore, in this research, the research team decided to use some outstanding DL models to train the state recognition of tomatoes based on image processing. On the other hand, these models were published mainly from 2014 to 2021 with the latest model being applied EfficientNet version 2. The parameters of the models are listed in TABLE 2. On the other hand, before the training process takes place, these models will be set up with a fixed number of hyperparameters to ensure the efficiency between the models. At the same time, the model will be fine-tuned, retaining the entire structure of the original layer and changing the output with Global Average Pooling, Dense and Dropout layers to limit overfitting and achieve fast results. The fixed parameters are shown in TABLE 3, and the model's basic structure is illustrated in FIGURE. 5.

E. GRAD-CAM APPROACH

In this research, the training phase and classification of objects in the image are used the entire DL models. As discussed in the previous section, the DL model is considered a "black box" model because it is impossible to dig into how the network structure works in image classification. Therefore, the research team uses Grad-CAM [3] to visualize and understand how a model predicts objects or features in an image based on the classification results of DL models. This technique is based on computing the gradient of the last convolution layer according to the feature maps of the model based on the specified class. Then, the model averages the gradient values to obtain a matrix representing the interest of the model for each position of that class feature in the image. In three steps, compute the gradient of output before softmax activation concerning the last convolution activation layer of the model, Average Alpha, by Averaging the gradients, performing a weighted combination of activation maps and following by ReLU to get the matrix from the equation (1). The resulting matrix will be combined with the original image data to represent the features of the image prominently. Illustrate the matrix constituting the heatmap with Viridis-toned feature regions in FIGURE. 6 and FIGURE 7.

$$L^c_{Grad-CAM} = ReLU \underbrace{\left(\sum_k \alpha^c_k A^k\right)}_{linearcombination} \tag{1}$$

F. ASSESSMENT MEASURES

In Stage 1, the research team used basic metrics to evaluate the image recognition models of the DL models. First, Precision and Recall were used to evaluate the model's correct predictions based on the positive prediction rate according to (2) and (3). However, we used the F1-Score metric to evaluate the completeness and reliability of the model while limiting the correlation between the two Precision and Recall parameters and providing the most objective evaluation among the models (4). In addition, accuracy was used to measure the effectiveness of an image classification model. It is calculated as the ratio of correctly predicted points to the total number of points in the test dataset (5). Finally, during the prediction process on the training and validation datasets, we also used the loss function parameter to evaluate(6).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = 2\frac{Precision * Recall}{Precision + Recall} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{5}$$

$$Loss = -\sum_{i=1}^{outputsize} y_i \cdot log\widehat{y_i} \tag{6}$$

In which:

- True Positive (TP): the number of points of the positive class that are correctly classified as positive.
- True Negative (TN): the number of points of the negative class that are correctly classified as negative.
- False Positive (FP): the number of points of the negative class that were mistakenly classified as positive.

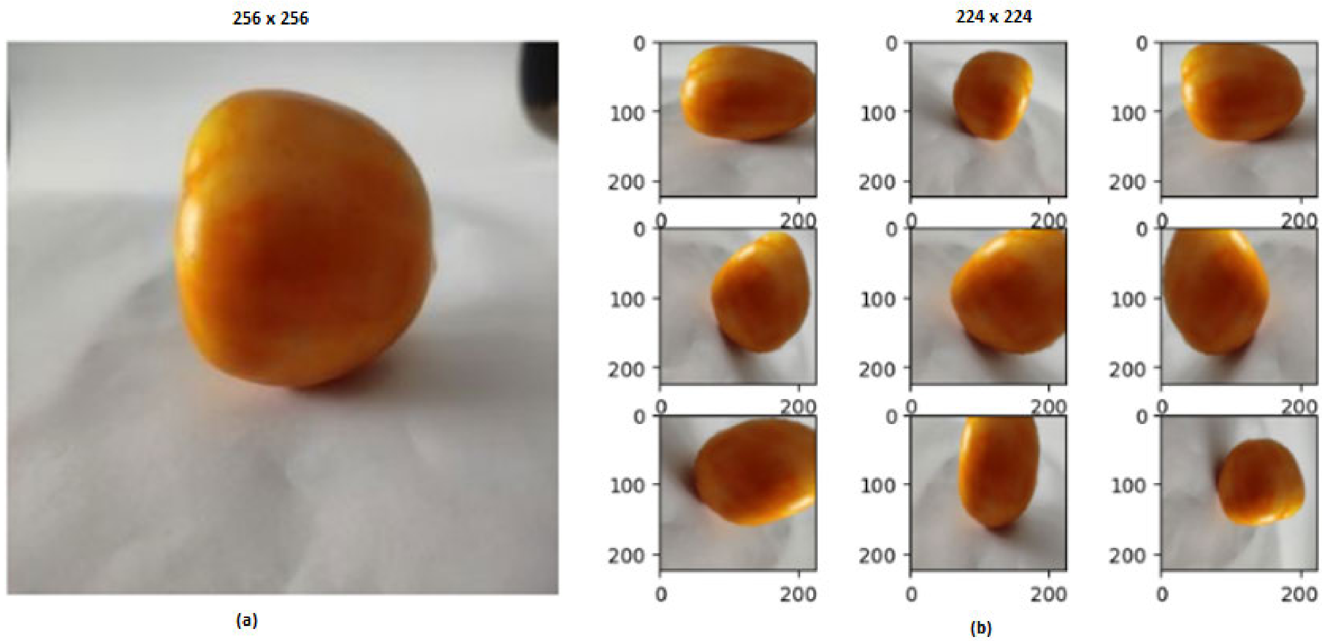


FIGURE 3. Illustrate the image in the dataset when the image is resized and argumentation is performed. (a) The original image in the dataset is 256 × 256 pixels. (b) Images are generated from image rotations and reduced to 224 × 224 pixels.

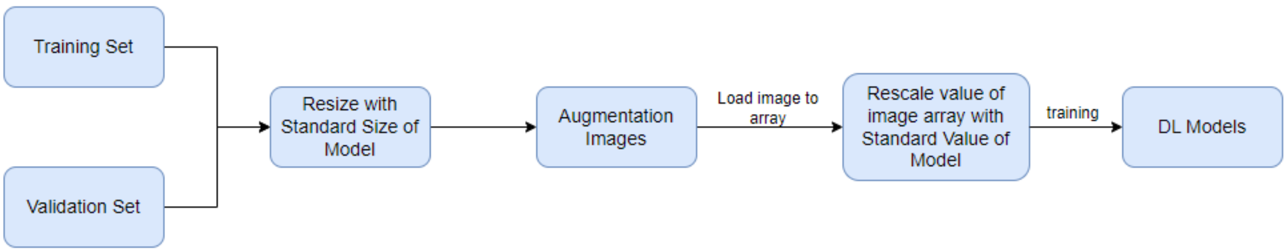


FIGURE 4. Illustrate the process of image data processing.

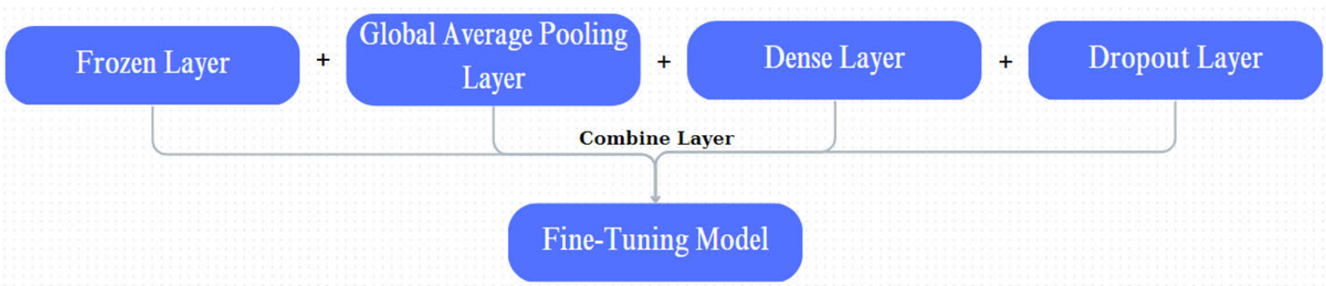


FIGURE 5. Illustrate the basic network structure of each model after applying Fine-tuning technique. With Frozen layer is the entire structure of the original model except for the output layer.

- False Negative (FN): the number of points of the positive class that were mistakenly classified as negative.

IV. RESULTS

A. COMPARISION PERFORMANCE OF DL MODELS BASED-ON RESULT OF TRADITIONAL METRICS

The results shown in FIGURE. 8 show that most of the models have very high accuracy, with more than 90% on both training and testing sets with epoch = 100. The accuracy of MobileNet and ResNet50 achieved almost 99% with the training set, and the accuracy of MobileNet and RegNetX002 reached 98%. However, the VGG16 model achieved relatively lower results than other models in about 100 epochs, with the accuracy achieved on the raining and validation set of 87% and 85%, respectively. This shows the effectiveness of the predictive model on the current dataset.

For the Loss Function parameter, the models all achieved similar results to the accuracy parameter. FIGURE. 9 shows the training and validation process between models quite well. When the parameters reach well at the threshold epoch = 100.Besides, models with the lowest loss function value in both training and validation progression are InceptionV3 and

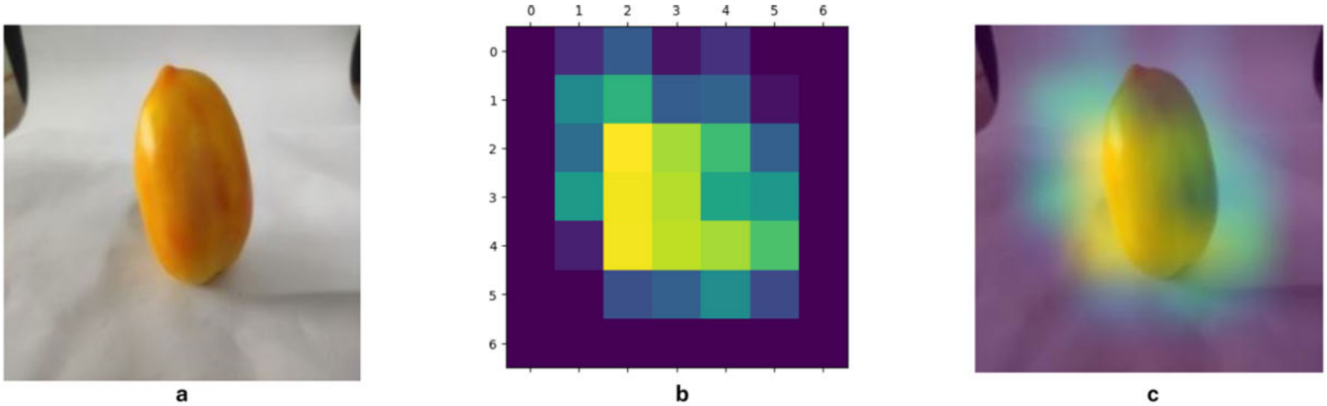


FIGURE 6. The image data is applied Grad-CAM to identify the display area of the feature on the Ripe label feature. (a) Original input image. (b) The matrix is made up of the gradient of the Grad-CAM algorithm. (c) Illustrate the Saliency from the matrix (b) onto the original image.

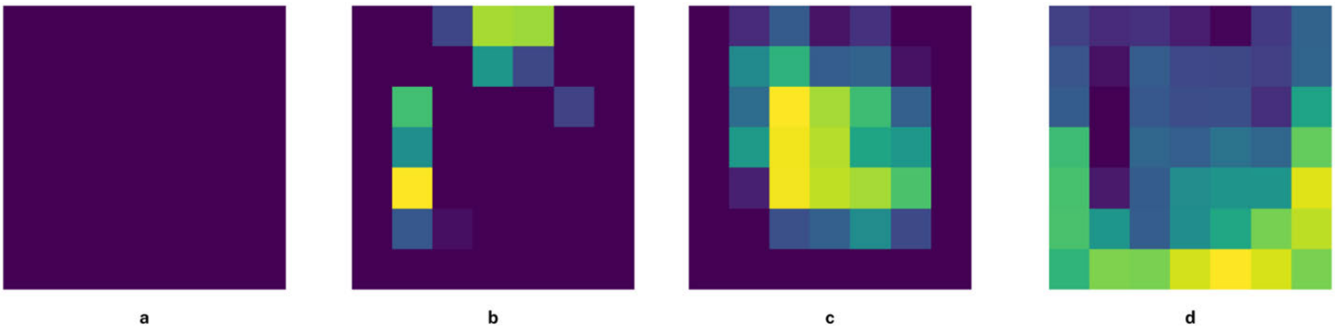


FIGURE 7. The feature matrix is calculated from the weight of each class based in FIGURE. 6a: (a) Damaged label features. (b) Characteristic of the Old label. (c) Characteristic of the Ripe label. (d) Features of the Unripe label.

Resnet50. In addition, during training, the chart shows that the two models learn features of the dataset well and are not biased and affected by different batch sizes of images. The valued line of the Accuracy or Loss Function graph does not change direction too profoundly after each epoch.

For accuracy evaluation, precision, recall and F1-Score parameters are shown in TABLE 4. The obtained results are relatively good on the unseen data dataset. At the same time, the remaining models achieved more than 90% of all parameters, and the lowest result is 91% with InceptionV3 model. The four models showing the most efficient, including ResNet50, EfficientNetV2, DenseNet201 and MobileNet model, have achieved more than 97% measurement parameter values. The traditional metrics for the image recognition model prove that MobileNet has the best results on all 3 datasets

B. LOCAL EXPLAINABILITY MODEL AND COMPARISON WITH GRAD-CAM

The research's main objective is to evaluate the reliability of the recognition model. Therefore, the research team uses Grad-CAM on feature-by-feature random image data to create 4 feature matrices, as shown in FIGURE. 7. Based on this result, the research team displays the Saliency Map on the original image. The result is an image of a damaged tomato with a damaged feature detected by Grad-CAM (FIGURE. 10). Based on each trained model, the models learn the features of the damaged area, except for the DenseNet201 and NasNetMobile models, which recognize the background and the intact part of the tomato. This result shows that these models achieve certain reliability for the feature of the Damage label. However, for the feature of Old label, Grad-CAM displays untrusted features because the regions identified by the DL are confused and coincide with the damaged area. The results of MobileNet, ResNet50, Xception, InceptionV3 and EfficientNetV2 models do not detect image features on the damaged fruit region (FIGURE. 10).

In the feature of the Ripe label, these models were correct when identifying most of the background and border regions of the tomato, similar to the Unripe labels(FIGURE. 11). The image illustrated shows that the best-performing models are Xception, InceptionV3, MobileNet and EfficientNetV2. In which the features modelled EfficientNetV2 and Xception are modelled accurately and are not affected much by other features on the image. Thereby, it can be seen that, although the MobileNet model achieves the highest results, the most reliable models are EfficientNetV2 and Xception

In TABLE 5, the research performs statistics on each image based on correctly predicting the features after using DL with the region displayed from Grad-CAM called Match Ratio

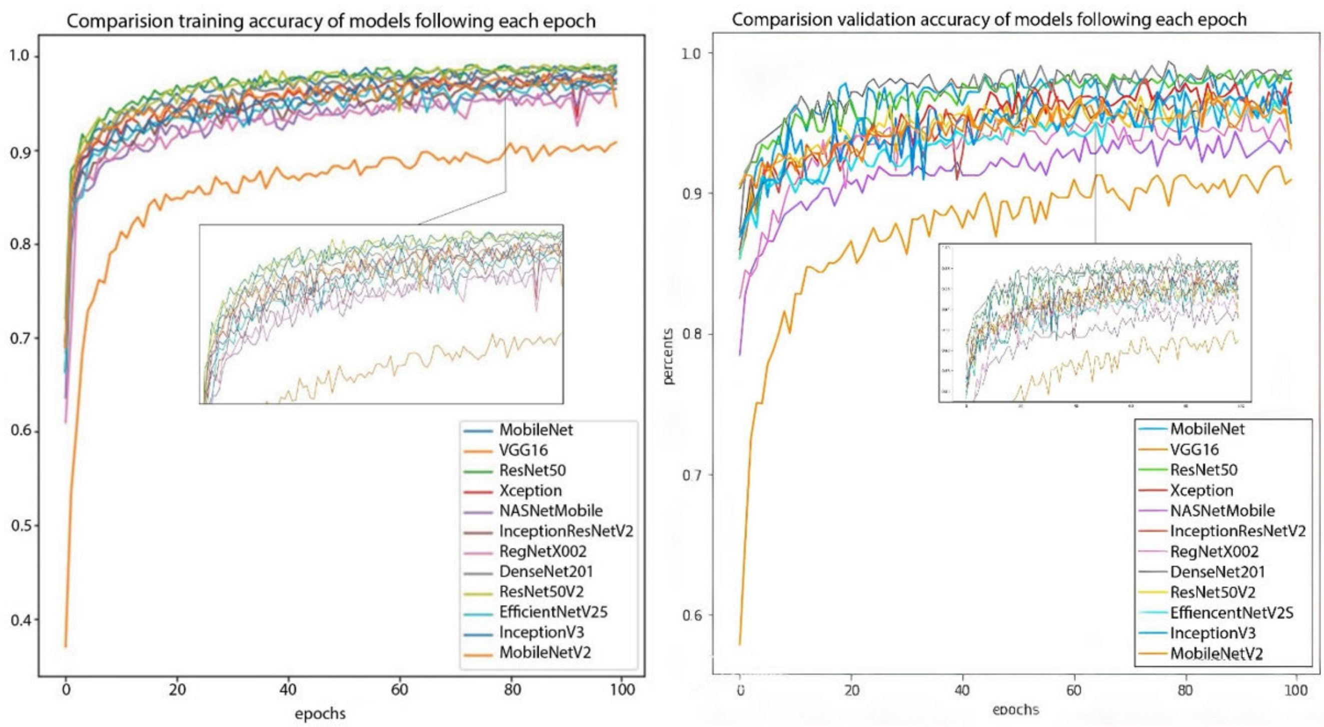


FIGURE 8. Comparative accuracy chart of DL models based on training and validation datasets.

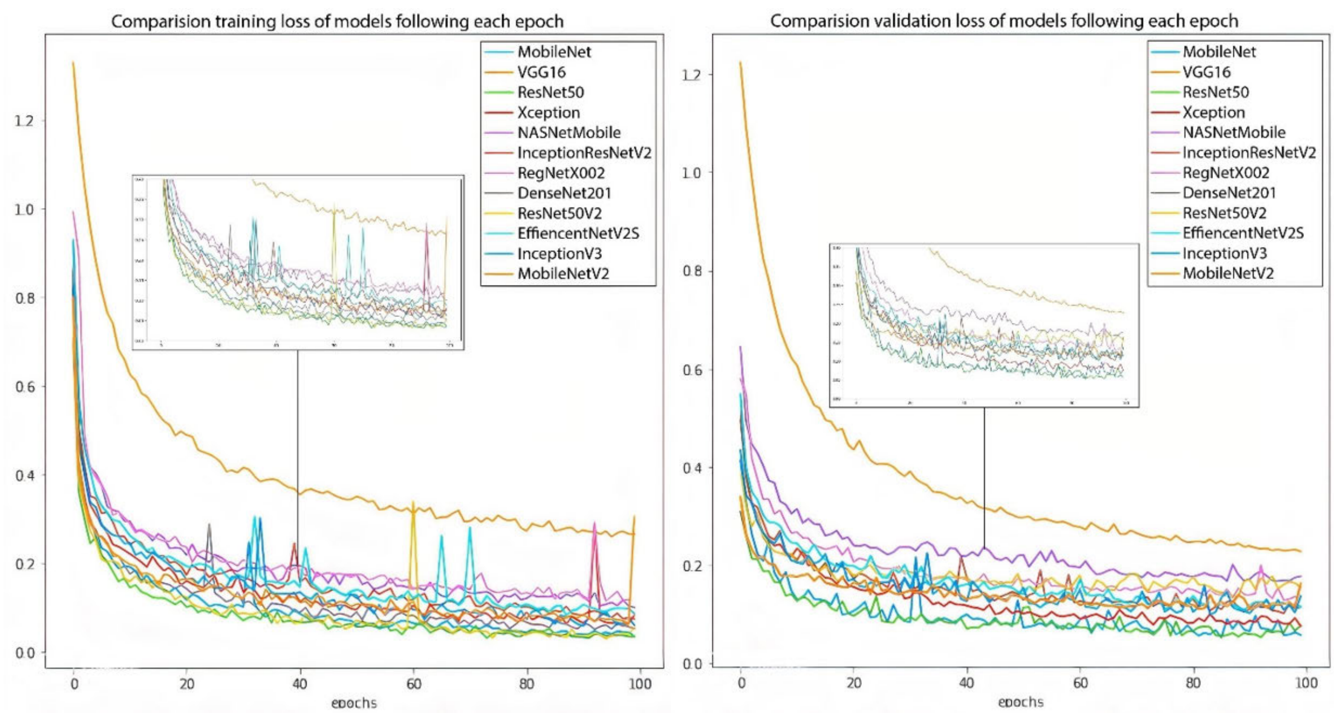


FIGURE 9. Comparison chart of loss function values of DL models based on Training and Validation datasets.

(MR). In case the data representation matches, the prediction is considered correct, else is false. With the statistical value in TABLE 5, we can see that the MR rate of MobileNetV2 is relatively high when the exact weight is from 97.09% to 100%. Meanwhile, MobileNet predicts the indexes to be relatively high, but the results show that choosing the right feature area still needs to be improved. However, in statistics, it can be seen that the Old and Damaged class does not have several 100% MR rates, and the average MR is smaller than Unripe and Ripe, which shows this feature is hard for the model can be learned clearly than other feature labels. For Ripe and Unripe, the MR ratio is 100% achieved by some

TABLE 4. Result table of the fine-tuning model on testing set.

DL model	Accuracy(%)	Precision(%)	Recall(%)	F1-Score
VGG16 (VG16)	63.00	75.8	63.0	58.00
ResNet50 (RES1)	97.32	97.39	97.34	97.34
ResNet50V2 (RES2)	95.16	95.00	94.84	94.88
Xception (XCEP)	97.24	96.88	96.84	96.83
EfficientNetV2 (EFN2)	97.11	96.67	96.58	96.62
InceptionV3 (INV3)	91.12	91.32	93.32	91.11
DenseNet201(D201)	97.10	96.71	96.66	96.66
MobileNetV2 (MON2)	96.00	96.10	95.57	95.73
MobileNet(MON1)	**98.00**	**97.60**	**97.50**	**97.50**
NasNetMobile(NANM)	95.40	95.38	95.00	95.15
RegNetX002(REN2)	96.41	96.00	95.85	95.91
InceptionResNetV2(IRN2)	96.60	95.86	95.61	95.71

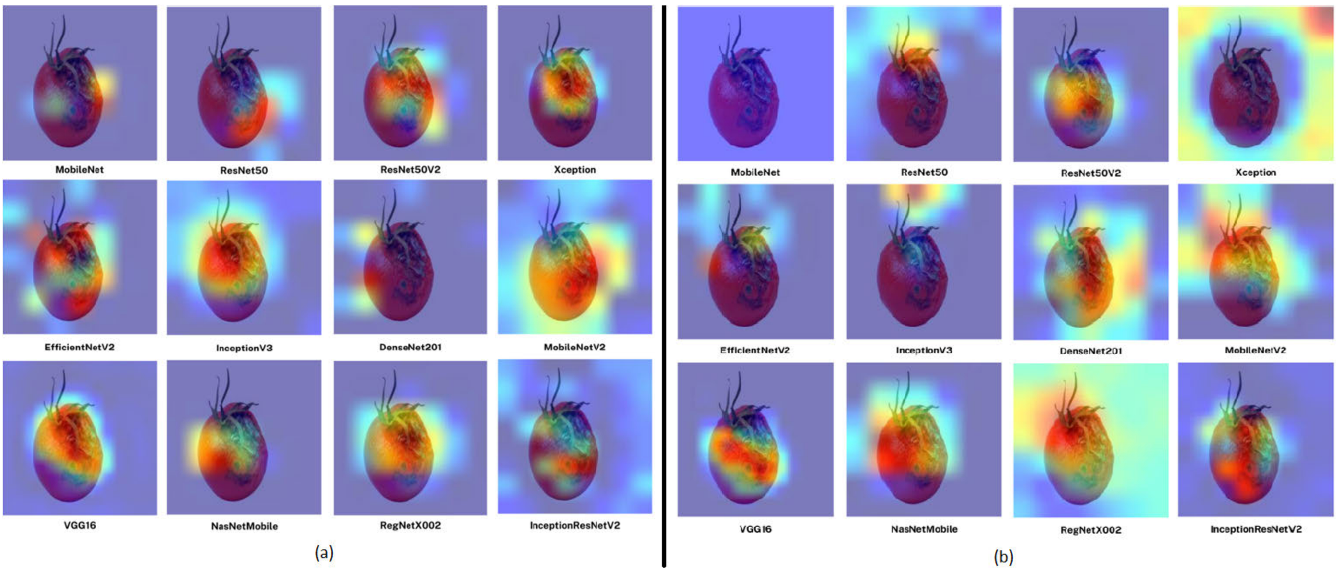


FIGURE 10. The Saliency Map based on the Grad-CAM matrix is displayed on the original image of the DL models: (a) Damaged feature, (b) Old feature.

TABLE 5. Statistical table of the ratio of displaying correctly predicted feature area when using grad-cam on each class and DL models (%).

Class	MON1	VG16	RES1	XCEP	NANM	IRN2	REN2	D201	RES2	EFN2	INV3	MON2
Damage	53.26	97.85	79.35	81.72	77.42	100.00	92.47	47.31	96.77	97.85	97.85	100.00
Old	91.26	66.99	85.44	98.06	93.20	97.09	100.00	54.37	99.03	98.06	94.17	97.09
Ripe	80.00	73.04	56.52	99.13	31.30	100.00	96.52	93.91	86.09	100.00	94.78	100.00
Unripe	100.00	67.71	98.96	95.83	47.92	98.96	96.88	97.92	97.92	98.96	98.96	100.00

models, Grad-CAM detects the features, but these features still match the feature matrices from the other labels.

V. DISCUSSION

The research proposes two stages in implementing the XAI model with Grad-CAM, showing relatively optimistic results in evaluating DL models in image classification, namely the evaluation of tomato fruit quality with images. In phase 1, the results show that the MobileNet model achieves the highest efficiency on the proposed dataset. This result is because the architecture of the MobileNet model uses special techniques such as depthwise convolution and pointwise convolution, which significantly reduces the number of parameters in the model, thereby reducing the computational complexity. Training speed is increased, and efficiency is demonstrated. However, in phase 2, Grad-CAM use shows that MobileNet is less reliable than EfficientNetV2, Xception and MobileNetV2 despite the higher recognition results. This can confirm that the accuracy assessment has the best results, but it is not necessarily the most complete and adequate model. In addition, using XAI to clearly explain what happens in the black box so that they become clearer is essential for making in-depth assessments and ensuring reliability in practice. FIGURE. 12 shows that with the InceptionResNetV2 model, although the testing indexes are over 95%, the learned features do not have certain reliability. From there, the model can misjudge in reality if the image is incomplete

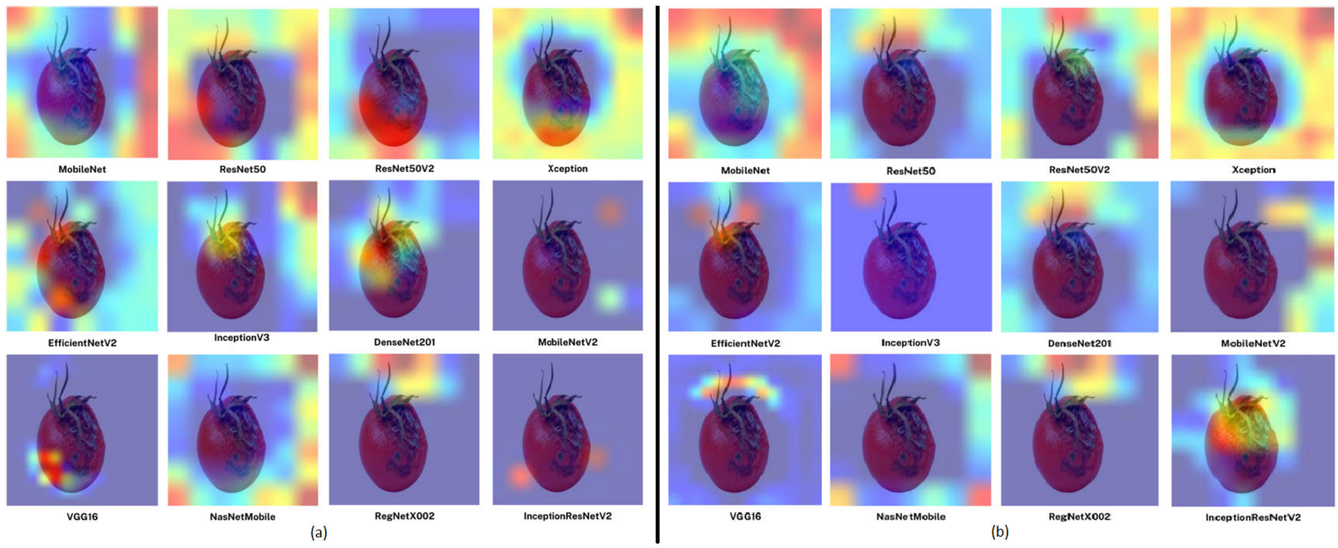


FIGURE 11. The Saliency Map based on the Grad-CAM matrix is displayed on the original image of the DL models: (a) Ripe feature, (b) Unripe feature.

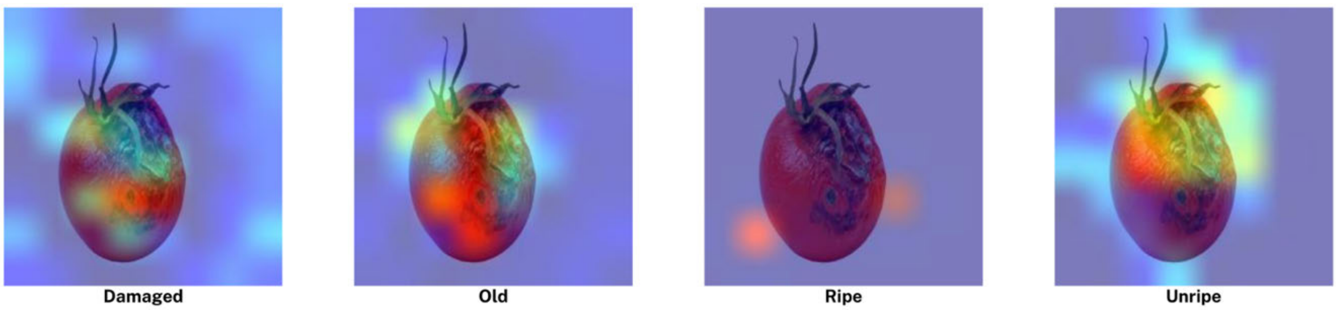


FIGURE 12. Saliency Map features 4 classes on Damaged image based on Grad-CAM matrix explained from InceptionResNetV2 model.

or the background becomes different. For example, removing the background and focusing only on the model object will misjudge from Damaged to Unripe because the image has only Unripe features. Through Grad-CAM, it is possible to compare the performance of the models clearly and ensure that the model learns essential features from the XAI based on which can improve the model and complete the dataset in a particular direction to ensure the model has higher reliability in the research.

VI. CONCLUSION

In this research, the research team compared the performance of DL models in the Image Classification task on the tomato images dataset. In the first stage, MobileNets demonstrated the highest efficiency in processing and grading tomatoes with 97.5% to 98% accuracy on the evaluation metrics and an average of 1% more than the models remaining. In the second stage, as mentioned above, the MobileNetsV2 model has the highest reliability through the Grad-CAM technique, with the proportion of vital features found by the model reaching 97% in the Damaged class and 100% in the Damaged class and 18.14% more than MobileNet model in each class. These results show that EfficientNetV2, Xception and MobileNetV2 models have higher reliability than MobileNet, although MobileNet is more efficient in the first stage of training. The results show that MobileNetV2 is the most effective of the models used for evaluation when the data is considered by segmentation. However, the reliability of a model is not based solely on the results of the test set but should be further evaluated using XAI to explain the model's decisions. The research team encourages using XAI in research and practice to ensure the model's reliability, especially in agriculture. At the same time, the study also proves that using XAI or Grad-CAM to evaluate the entire black box model to develop the most reliable model is essential. This proposal will help improve the model and complete the dataset in a particular direction for a more reliable model in future research.

VII. FUTURE WORK

Although this research has contributed valuable results on the effectiveness of DL models in Image Classification, as well as partially solved the problem of tomato fruit state recognition in the agricultural field. It also gives the importance of XAI to Computer Vision in completing and evaluating the black box model. In the future, a further direction could be to expand the research's object by increasing the dataset's specific complexity. This work will pose many challenges regarding the speed and accuracy of DL models. In addition, when developing a DL model or system, finding ways to optimize its performance is necessary. In addition, XAI systems need to research methods of evaluating Grad-CAM algorithm results, such as model-based evaluations instead of relying solely on human-based evaluations, to move from local explainability to global explainability is also a potential future research direction. This method will help improve the reliability of DL models in real-world tasks across the entire dataset.

REFERENCES

[1] L. Gianfagna and and A. Di Cecco, *Explainable AI With Python*. Cham, Switzerland: Springer, 2021.

[2] Y. Suryawanshi, K. Patil, and P. Chumchu, "VegNet: Dataset of vegetable quality images for machine learning applications," *Data Brief*, vol. 45, Dec. 2022, Art. no. 108657.

[3] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," 2016, *arXiv:1610.02391*.

[4] H. R. Boveiri, R. Khayami, R. Javidan, and A. Mehdizadeh, "Medical image registration using deep neural networks: A comprehensive review," *Comput. Electr. Eng.*, vol. 87, Oct. 2020, Art. no. 106767.

[5] S. S. Han, I. Park, S. Eun Chang, W. Lim, M. S. Kim, G. H. Park, J. B. Chae, C. H. Huh, and J.-I. Na, "Augmented intelligence dermatology: Deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders," *J. Investigative Dermatol.*, vol. 140, no. 9, pp. 1753–1761, Sep. 2020.

[6] A. T. Sahlol, M. A. Elaziz, A. T. Jamal, R. Damasevicius, and O. F. Hassan, "A novel method for detection of tuberculosis in chest radiographs using artificial ecosystem-based optimisation of deep neural network features," *Symmetry*, vol. 12, no. 7, p. 1146, Jul. 2020.

[7] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, and G. Srivastava, "Deep neural networks to predict diabetic retinopathy," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 5, pp. 5407–5420, Apr. 2020.

[8] L.-D. Quach, N. P. Quoc, N. H. Thi, D. C. Tran, and M. F. Hassan, "Using SURF to improve ResNet-50 model for poultry disease recognition algorithm," in *Proc. Int. Conf. Comput. Intell. (ICCI)*, Oct. 2020, pp. 317–321.

[9] L.-D. Quach, N. P. Quoc, N. H. Thi, N. L. Thanh, X. T. T. Thanh, D. C. Tran, and M. F. Hassan, "An AI-based chicken disease management system," in *Proc. Int. Conf. Artif. Intell. Smart Community*. Singapore: Springer, 2022, pp. 707–716.

[10] L.-D. Quach, K. N. Quoc, A. N. Quynh, and H. T. Ngoc, "Evaluation of the efficiency of the optimization algorithms for transfer learning on the rice leaf disease dataset," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 10, pp. 83–91, 2022.

[11] N. Duong-Trung, L.-D. Quach, and C.-N. Nguyen, "Towards classification of shrimp diseases using transferred convolutional neural networks," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 5, no. 4, pp. 724–732, 2020.

[12] Y. Mintz and R. Brodie, "Introduction to artificial intelligence in medicine," *Minimally Invasive Therapy Allied Technol.*, vol. 28, pp. 73–81, Feb. 2019.

[13] M. O. Riedl, "Human-centered artificial intelligence and machine learning," *Hum. Behav. Emerg. Technol.*, vol. 1, no. 1, pp. 33–36, Jan. 2019.

[14] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1625–1634.

[15] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Berlin, Germany: Springer, 2019, pp. 5–22.

[16] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: An analytical review," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 11, no. 5, p. e1424, Jul. 2021.

[17] Z. Huang, X. Zhu, M. Ding, and X. Zhang, "Medical image classification using a light-weighted hybrid neural network based on PCANet and DenseNet," *IEEE Access*, vol. 8, pp. 24697–24712, 2020.

[18] M. M. Hassan, S. A. AlQahtani, A. Alelaiwi, and J. P. Papa, "Explaining COVID-19 diagnosis with Taylor decompositions," *Neural Comput. Appl.*, pp. 1–14, Nov. 2022.

[19] V. Jahmunah, E. Y. K. Ng, R.-S. Tan, S. L. Oh, and U. R. Acharya, "Explainable detection of myocardial infarction using deep learning models with grad-CAM technique on ECG signals," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105550.

[20] X. Wang, X. Liang, Z. Jiang, B. A. Nguchu, Y. Zhou, Y. Wang, H. Wang, Y. Li, Y. Zhu, F. Wu, J. Gao, and B. Qiu, "Decoding and mapping task states of the human brain via deep learning," *Hum. Brain Mapping*, vol. 41, no. 6, pp. 1505–1519, Apr. 2020.

[21] J. Streich, J. Romero, J. G. F. M. Gazolla, D. Kainer, A. Cliff, E. T. Prates, J. B. Brown, S. Khoury, G. A. Tuskan, M. Garvin, D. Jacobson, and A. L. Harfouche, "Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the united nations sustainable development goals?" *Current Opinion Biotechnol.*, vol. 61, pp. 217–225, Feb. 2020.

[22] C. Arvind, A. Totla, T. Jain, N. Sinha, R. Jyothi, K. Aditya, M. Farhan, G. Sumukh, and G. Ak, "Deep learning based plant disease classification with explainable AI and mitigation recommendation," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2021, pp. 1–8.

[23] H. Singh, A. Roy, R. K. Setia, and B. Pateriya, "Estimation of nitrogen content in wheat from proximal hyperspectral data using machine learning and explainable artificial intelligence (XAI) approach," *Model. Earth Syst. Environ.*, vol. 8, no. 2, pp. 2505–2511, Jul. 2021.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2016, pp. 630–645.

[27] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[28] M. Tan and Q. V. Le, "EfficientNetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[30] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 4700–4708.

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[32] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[33] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.

[34] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10425–10433.

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, Feb. 2017, pp. 1–7.

[36] C. I. Nwakanma, L. A. C. Ahakonye, J. N. Njoku, J. C. Odirichukwu, S. A. Okolie, C. Uzondu, C. C. N. Nweke, and D.-S. Kim, "Explainable artificial intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: A review," *Appl. Sci.*, vol. 13, no. 3, p. 1252, Jan. 2023.

[37] D. Saraswat, P. Bhattacharya, A. Verma, V. K. Prasad, S. Tanwar, G. Sharma, P. N. Bokoro, and R. Sharma, "Explainable AI for healthcare 5.0: Opportunities and challenges," *IEEE Access*, vol. 10, pp. 84486–84517, 2022.

[38] D. Vale, A. El-Sharif, and M. Ali, "Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law," *AI Ethics*, vol. 2, no. 4, pp. 815–826, Mar. 2022.

[39] D. V. Kute, B. Pradhan, N. Shukla, and A. Alamri, "Deep learning and explainable artificial intelligence techniques applied for detecting money laundering—A critical review," *IEEE Access*, vol. 9, pp. 82300–82317, 2021.

[40] S. Sachan, J.-B. Yang, D.-L. Xu, D. E. Benavides, and Y. Li, "An explainable AI decision-support-system to automate loan underwriting," *Exp. Syst. Appl.*, vol. 144, Apr. 2020, Art. no. 113100.

[41] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019.

[42] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[43] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun, "Deep learning for image-based cancer detection and diagnosis—A survey," *Pattern Recognit.*, vol. 83, pp. 134–149, Nov. 2018.

[44] R. Brehar, D.-A. Mitrea, F. Vancea, T. Marita, S. Nedevschi, M. Lupsor-Platon, M. Rotaru, and R. I. Badea, "Comparison of deep-learning and conventional machine-learning methods for the automatic recognition of the hepatocellular carcinoma areas from ultrasound images," *Sensors*, vol. 20, no. 11, p. 3085, May 2020.

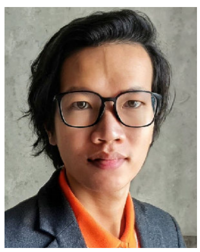
LUYL-DA QUACH received the B.S. degree in information technology from Tay Do University, Vietnam, and the master's degree in information systems from Can Tho University, Vietnam, in 2016. He is currently pursuing the Ph.D. degree with the Ho Chi Minh City University of Technology, Vietnam.

Since 2018, he has been a Lecturer and a Researcher with the Department of Software Engineering, FPT University, Can Tho, Vietnam. His field of interest is applying information technology and artificial intelligence to solve a specific problem. His current research interests include image processing, intelligent systems, natural language processing, and explainable AI. He is a Reviewer for several specialized journals, such as *Expert Systems With Applications*, *Information Sciences*, and several international conferences in Vietnam, Malaysia, and Bahrain.

KHANG NGUYEN QUOC is currently pursuing the B.S. degree in software engineering with FPT University, Vietnam. His current research interests include computer vision, especially in processing images and video for the detection of objects and analysis of data.

ANH NGUYEN QUYNH is currently pursuing the B.S. degree in software engineering with FPT University, Can Tho, Vietnam. Her current research interests include applying deep learning to develop agriculture, especially in diagnosing and classifying diseases to increase the productivity of shrimp and rice.

NGUYEN THAI-NGHE received the B.S. degree in informatics from Cantho University (CTU), the M.S. degree in information management from the Asian Institute of Technology, Thailand, and the Ph.D. degree in computer science from the University of Hildesheim, Germany. From 2009 to 2012, he got a scholarship from WorldBank-CTU to do Ph.D. degree. He is currently an Associate Professor with the Department of Information Systems, Cantho University. He is a PC Member/Reviewer of several international conferences and journals, such as *Future Data and Security Engineering* (Springer), IEEE ACOMP, IEEE KSE, Springer ACIIDS, *PLOS One*, and *Advances in Science, Technology and Engineering Systems Journal*.



TRI GIA NGUYEN (Senior Member, IEEE) is currently the Academic Director of FPT University, Quy Nhon AI Campus, Vietnam. His current research interests include the Internet of Things, sensor networks, wireless communications, wireless energy harvesting networks, mobile computing, edge computing, software-defined networking, network functions virtualizations, and network security.