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

Leveraging Multi-Round Learning and Noisy Labeled Images from Online Sources for Durian Leaf Disease and Pest Classification

Sasin Janpuangtong[1] and Kritwara Rattanaopas[2]

ABSTRACT

Durian has recently become a major agricultural export commodity for Southeast Asian countries. However, this plant is vulnerable to various diseases and pests, which are usually considered the main cause of poor yields and low-quality crops; leading to a huge economic loss. This work focuses on leaf diseases and pests, as their symptoms can be easily detected visually, but it is still challenging to correctly diagnose the problems. To address this difficulty, this work makes use of deep learning techniques to classify a given photo to a corresponding class of disease or pest. However, building a high-performance deep neural network model requires a substantial amount of ground-truth photos of diseases and pests on durian leaves, which are difficult and expensive to acquire. To overcome this challenge, we propose enriching the limited number of expert-labeled images with abundantly available, noisily labeled images collected from the Internet. A sample selection framework is introduced to choose noisy images for augmenting a current training set, which will be used to build a new classifier in the next learning round. We found that such a multi-round learning scheme, in which noisy photos are intuitively selected, provides complementary information to the limited ground truth, thereby enhancing the prediction accuracy on unseen examples of a classifier being built by 20% at a particular learning round.

Article information:

Keywords: Weak Supervision, Multi-round Learning, Image Classification, Durian Leaf Diseases and Pests

Article history:
Received: August 25, 2024
Revised: November 28, 2024
Accepted: January 2, 2025
Published: January 18, 2025 (Online)

DOI: 10.37936/ecti-cit.2025191.258069

1. INTRODUCTION

Durian is a tropical fruit well known for its peculiar look and smell, but it is gaining global popularity due to its distinctive sweet and creamy taste. The "King of Fruits, which is cultivated largely in Southeast Asia, has become one of the important exports that generates huge revenues for countries in this region. However, there are many tropical diseases and pests that are the main threats to this crop, severely affecting the quality and quantity of the yields. A fast and accurate classification of these diseases and pests plays a key role in protecting crops against severe and uncontrollable damage [1]. This work focuses on two diseases (algal spot and anthracnose) and two pests (mealybugs[3] and pit scale[4]) of Durian leaves, which are commonly found in tropical farming areas. Although their symptoms may be visually detectable on the leaves, it is challenging for farmers to identify the problem themselves, as they may not have enough knowledge or confuse one with others (see Fig. 1). The conventional method used by agricultural experts, in which symptoms must be carefully examined or tested in a laboratory, is extremely laborious and costly [2]. To address such difficulties, a new approach using deep learning techniques is introduced [3, 4] to learn a neural network model, especially convolutional neural networks (CNN), which can correctly classify a given Durian leaf photo into a correct class of disease or pest.

[1]The author is with the Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Thailand, E-mail: sasin.ja@cmu.ac.th
[2]The author is with the Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University, Thailand, E-mail: kritwara.r@psu.ac.th
[2]Corresponding author: kritwara.r@psu.ac.th
[3]*Allocaridala maleyensis Crawford*
[4]*Asterolecanium ungulatum Russell*

(a) Algal spot (b) Mealybugs (c) Pit scale (d) Anthracnose

Fig.1: *The examples of human-labeled images of four Durian leaf diseases and pests that we aim to classify in this work. These photos were taken as farmers were supposed to. So, Durian leaves in each photo are under heterogeneous conditions in terms of the background, illumination direction, leaves and symptoms position, and so on.*

(a) Algal spot (b) Mealybugs (c) Pit scale (d) Anthracnose

Fig.2: *An example of a noisy, labeled image for each class is obtained by submitting a query associated with that class to a web search engine. The retrieved images are often vastly different from the corresponding ground truths. These discrepancies arise not only from variations in image quality, lighting, and focal points, but also from the inherent differences in the appearance of the diseases and pests on the leaves themselves. It is important to note that some of the retrieved images do not even depict Durian leaves. These factors collectively make the labeling task particularly challenging, and even domain experts may struggle to differentiate between similar diseases or pests. Despite the incorrect labels, the large volume of such images available from various online sources presents an opportunity to intuitively enhance the performance of deep neural networks.*

Although deep learning has achieved outstanding success in various fields and tasks [5–11], this learning technique requires a large amount of cleanly labeled data (*i.e.*, ground truth) to produce a high-performance neural network. Ground truth examples may be expensive or impossible to collect in real-world circumstances for certain domains and tasks [12, 13], which include Durian leaf disease and pest classification, since it is a local crop and only a handful of domain experts are available. Various approaches have been introduced in recent years to circumvent such a difficulty, so high-quality models can be built from the insufficient number of labeled images. The simplest is augmentation, in which additional training images are constructed by applying simple transformations to the given ground truth [14]. Another common approach is transfer learning [15], which makes use of pre-trained weights learned from a large set of images for a certain task and then fine-tunes those weights further regarding the target task. Since the model does not train from scratch, a small number of ground-truth images may be sufficient to produce a good model; however, the effectiveness of this approach depends on the similarity of the source and target tasks.

Semi-supervised deep learning [16] trains deep neural networks from a combination of a small amount of human-labeled data and a large amount of unlabeled data [17, 18]. In addition, learning techniques that build deep neural networks with good performance from noisy labeled images have been increasingly developed to cope with the scarcity of ground truth. [16, 19–21] Although their labels may be falsely assigned by different forms of mistake, as shown in Fig. 2, such images are abundantly available on the Internet and have proved to be valuable in improving the generalization of a model being built [16, 19–21]. This work proposes a learning framework, which adopts the concept of noisy sample selection [21, 22], to incrementally augment a given set of small ground truths with a selected set of noisy labeled Durian leaf photos retrieved from the Internet via a web search

engine. There are two main advantages of choosing noisy sample selection as an underlying technique: It does not rely on any particular assumption on the noise of the label and is sufficiently general to train various CNN architectures.

Note that our ultimate goal is to develop a mobile application that accurately classifies a given photo taken by a farmer to a corresponding class of Durian disease or pest, thereby enabling farmers to identify the problem themselves without any help from the domain experts. Toward this end, the present work mainly focuses on the question of how to build a high-performance classifier from the limited number of ground truths. The main contributions of this work are as follows: We introduce a multi-round learning framework that leverages online, noisy, labeled images retrieved using a web search engine to enhance the task of identifying plant diseases and pests on Durian leaves. To the best of our knowledge, this work is the first one that aims to make use of noisy labeled images from heterogeneous sources on the Internet to learn a classifier for this particular domain. Also, our work does not impose any assumptions or constraints over the images, such as resolution, quality, lighting condition, coloring, or the ratio of an interesting object used to train a classifier, which is different from existing works. Posing such restrictions on photos to be taken is not practical and does not match our main objective, which is that the mobile application should be easy to use for farmers who may not have been exposed to new technology or have a background in plant diagnostics.

2. BACKGROUND

In this section, we review some groundwork on deep learning with noisy labels. Research focusing on applying deep learning to the task of identifying leaf diseases is discussed. Several existing efforts are also emphasized on the classification of Durian leaf disease.

2.1 Deep learning with noisy labels

Using noisy labeled data as an additional source of information to supplement the limited number of ground truth labels has been widely studied in recent years. However, the performance of deep learning models is significantly degraded when training with noisy labels [16, 23]. As a result, substantial efforts have been made to improve the robustness of deep neural networks against noisy labels [19].

One popular approach to mitigate the impact of noisy labels is loss adjustment, which reduces the negative effects by modifying the loss associated with all training examples before updating the network weights [19]. Techniques related to this approach can be grouped into four general categories based on their adjustment philosophy. The first category, loss correction, works similarly to the noise adaptation layer described earlier [24, 25]. Loss re-weighting assigns lower weights to examples with incorrect labels and higher weights to those with true labels [26, 27]. The main challenge in implementing this approach is constructing an appropriate weighting function tailored to the specific type of noise. The third category, label refurbishment, adjusts the loss by replacing noisy labels with refurbished labels obtained through a convex combination of noisy and predicted labels [19]. Methods in this category include bootstrapping [28], D2L [29], SELFIE [21], and SEAL [30]. Our approach partially aligns with the concept of label refurbishment, where noisy labels are explicitly replaced with more reliable labels derived from the agreement between the original labels and those predicted by the deep neural network features.

Meta-learning has recently become an important topic in machine learning and has been applied to address noisy labels [31, 32]. The key idea behind meta-learning is to perform learning at a higher level than conventional learning, enabling the development of data-agnostic and noise-type-agnostic strategies for real-world applications. However, a major limitation of this approach is that unbiased and clean validation data, which may not be available in practice, are typically required to achieve the desired outcomes [19].

Our framework primarily draws from the sample selection approach, which focuses on identifying true-labeled examples from noisy ones through multi-network or multi-round learning [19]. Collaborative learning and co-training are common techniques for multi-network training [33–35]. Without the need for additional deep neural networks, multi-round learning refines the selection of clean examples through multiple rounds of training [36, 37]. Although this approach can train a model without supervision and is effective against heavy noise [19, 38], it has the drawback of a linear increase in computational cost as the number of training rounds grows [19].

An alternative promising direction is the hybrid approach, which combines a specific sample selection strategy with a semi-supervised learning technique. In this approach, selected examples are treated as clean labeled data, while the remaining examples are treated as unlabeled [39, 40]. An example of a hybrid approach similar to our framework is SELF [22], which employs semi-supervised learning to progressively filter out examples whose ensemble predictions, generated by the mean-teacher model, do not align with their annotated labels. Rather than employing the self-ensemble technique to construct the average model (i.e., mean-teacher), we propose using the self-learning technique [38] to iteratively improve a single model, incorporating additional examples whose labels are corrected in subsequent training rounds.

2.2 Identifying leaf diseases using deep learning

The recent breakthrough in applying deep learning techniques, especially CNNs, to the field of computer vision has played an important role in solving complex problems in various domains. Applications in agriculture also benefit from this prominent development, especially the detection and classification of crop diseases, as this new technology presents a plausible alternative to traditional practices [41]. The following are emerging efforts to detect and identify leaf diseases for different crops using CNNs [42]. Bedi and Gole [43] introduced a hybrid model based on a convolutional autoencoder network and CNN to detect bacterial spot disease on peach leaf images publicly available in $PlantVillage$[5] . A model based on the pretrained VGG-16 was developed to detect healthy and unhealthy tomato and grape leaves [44]. A conditional generative adversarial network (GAN) was then used to generate synthetic images of tomato leaves, and a separate model was trained on both real and generated images to classify them into ten categories of diseases [45]. While using a generative network to automatically create a set of synthetic images is a promising approach to address the limited availability of ground-truth images, training such a network is computationally expensive and requires a diverse set of training examples to generate realistic images.

Despite the extensive use of various deep learning techniques for leaf disease detection and classification [46, 47], these existing works generally assume that there is a sufficient number of correctly labeled images available to train a prospective model, which may not be valid in a real world situation. It should be noted that these studies focused merely on common crops, which are grown all over the world, so the information related to them and their leaf diseases is well known and widely shared. Furthermore, there are large expertly curated image datasets that are publicly contributed to such regular plants and their diseases, thereby building a high-performance leaf disease classifier for these crops can be achieved. As a local crop, some diseases and pests may be particular to Durian and its farming areas. Lack of domain experts, it is increasingly difficult to acquire a sufficient number of correctly labeled images to build a promising classifier for this crop. Consequently, making use of noisy labeled images that are abundantly available on the Internet could be a viable option to address the difficulty that occurred in local plants.

2.3 Durian leaf disease classification

Several efforts have been made to identify Durian leaf diseases, with many sharing the same goal as our work: to help farmers easily and accurately diagnose diseases on Durian leaves using mobile phones [3, 4]. In this context, MobileNet [48] has emerged as a popular deep neural network architecture, as it requires fewer parameters to be estimated while maintaining prediction performance comparable to larger architectures.

Although our work and existing approaches face the same limitation—only a small set of human-labeled examples—there is a key distinction in the images used to train the classifiers. In previous works, the images were captured under controlled conditions, where each photo contained a single leaf positioned centrally against a homogeneous background, with consistent angle, lighting, and focal points. These strict conditions allowed all features of the leaf to be clearly observed. However, requiring such controlled conditions for photos taken by farmers, who may have limited knowledge of new technology or plant diagnostics, is not practical and makes the application harder to use. In contrast, our work relaxes these constraints, allowing farmers to take photos of Durian leaves under varied conditions for disease identification.

Another major difference is in how we handle the limited number of labeled examples. While existing approaches rely on data augmentation techniques to artificially increase the number of training examples, our approach focuses on leveraging noisy labeled images retrieved from diverse online sources to augment the ground-truth dataset, rather than simply applying conventional data transformations. Although both techniques aim to increase the training dataset size, the former does not introduce new information, whereas our method has the potential to provide valuable new data that can help improve the model's ability to represent and classify leaf diseases and pests more effectively.

3. PROPOSED LEARNING FRAMEWORK

In this section, we describe the key ideas and processes that underlie the proposed framework.

3.1 Overview of the proposed framework

The terminology used in this work is based on the framework of deep self-learning from noisy labels [38]. Let S be a small ground-truth training dataset, where $S = \{X_s, Y_s\} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, containing n samples. Likewise, the noisy data set is denoted by $D = \{X_D, Y_D\} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, in which $N \gg n$. For each image x_i, its corresponding label is $yi \in \{1, 2, \cdots, K\}$, where K is the number of classes. However, the main difference between Y_S and Y_D is that the labels in Y_D are noisy. In addition, let T_r be a set of selected training examples for the r^{th} round of learning, that is, $\mathrm{T}_1 = \mathrm{S}$ for the first round of training. The new labels of the images in D are estimated

[5] https://www.kaggle.com/datasets/abdallahalidev/plantvillage-dataset

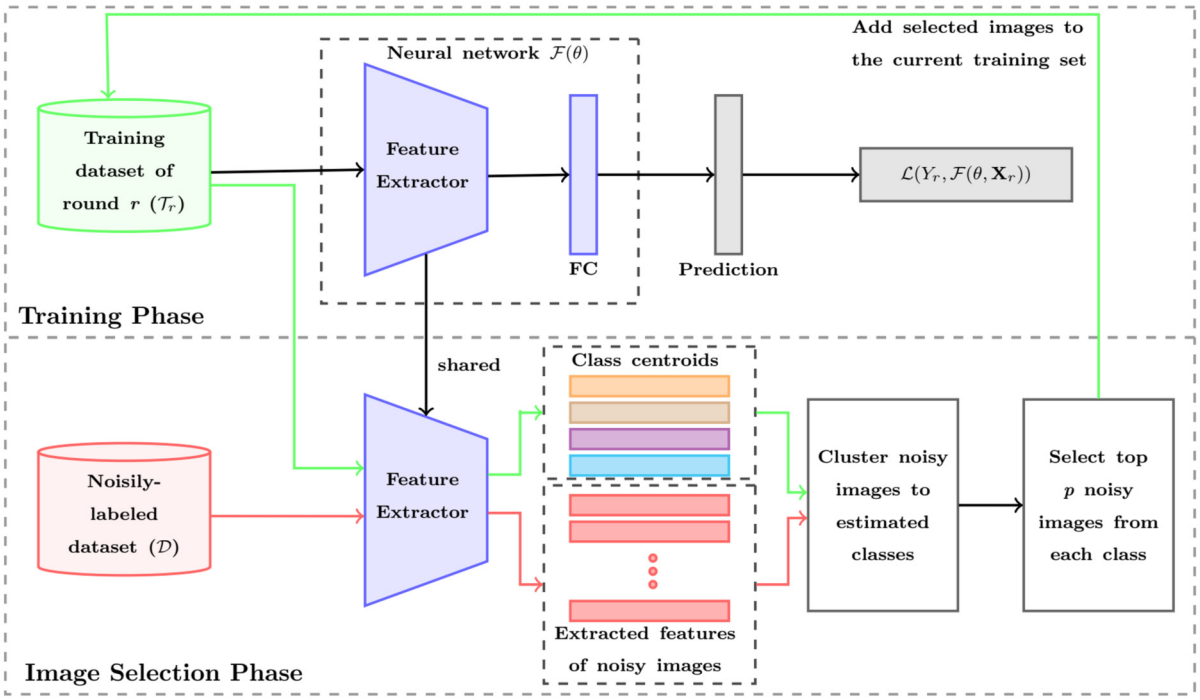

Fig.3: *The pipeline of the proposed iterative learning framework for noisy labeled datasets consists of two phases. The first phase, at the top, is the training phase, while the second phase, at the bottom, is the image selection phase. Notably, the convolutional part of the deep neural network can be shared across both phases, meaning that only a single neural network is built and evaluated. The green arrows indicate the flow of information corresponding to the training dataset, whereas the red arrows represent the flow of noisy labeled images.*

in a self-training manner, and p noisy labeled images whose new labels were estimated are then selected with respect to a particular metric introduced later in this section. The selected set of p images from D in r^{th} round, denoted by D_r, augments the current training set for the next learning round, so that $T_{r+1} = T_r \cup D_r$ and $\mathcal{D}, = \mathcal{D} \backslash D_r$. Given a training dataset $T_r = X_r, Y_r$, the following objective function is optimized at the r^{th} learning round.

$$\theta_r^* = \underset{\theta}{\operatorname{argmin}}\ \mathcal{L}(Y_r, \mathcal{F}(\theta, X_r)) \tag{1}$$

where $\mathcal{L}$ represents the empirical risk of cross-entropy loss and $\mathcal{F}(\theta, X_r)$ produces an estimated label for each image through a deep neural network $\mathcal{F}(\theta_{r-1})$ learned from the former round. To construct Dr, the new label $\hat{y}_i$ of an image $(x_i, y_i) \in \mathcal{D}$ is estimated with respect to the class representations learned from the previous round. The estimated label of a given image x_i is determined by

$$\hat{y}_i = \underset{c \in C}{\operatorname{argmax}}\ \delta(z_i, z_c) \tag{2}$$

where z_i is a set of extracted features from the convolutional part of a neural network, z_c is a set of features representing a centroid of class c, and δ is a similarity measurement. The representation of class c is estimated by

$$z_c = \frac{\sum_{i=1}^{|X_c|} z_i}{|X_c|} \tag{3}$$

where X_c is a set of training images that belongs to class c and $X_c \subset X_r$.

Although we believe that the estimated label $\hat{y}_i$ is more precise than y_i originally assigned to the image, the estimated class of many examples may still be highly likely false. To address this challenge, we developed a scoring scheme that takes into account both the original and estimated labels to which the top-scoring images are selected to be added to the current training set. Here, cosine similarity is used for δ, so that the corresponding score, denoted by ϕ_i of a given image $X_c \in D$ is in [0, 1.

$$\phi_i = \begin{cases} 1 & if\ y_i = \hat{y}_i \\ \alpha\delta(z_i, z_0) + (1-\alpha)\max\limits_{c \in C} \delta(z_i, z_c) & if\ y_i \neq \hat{y}_i \end{cases} \tag{4}$$

where the hyperparameter $\alpha \in (0, 1)$ imposes the confidence on the original label, and z_0 is the current centroid corresponding to the original label of a given image. Since we want to make sure that an image whose original label is similar to the estimated one is selected, as it shows the harmony between two distinct information sources, the score of such an image

becomes 1.

3.2 Framework architecture

The overall framework is illustrated in Figure Fig. 3, which contains two phases: training and noisy labeled image selection, working together in an iterative manner.

3.2.1 Training phase

In this phase, a deep neural network $\mathcal{F}$ with parameters θ is normally trained on the given training set. This phase is run repeatedly for a given number of rounds using a given set of ground truths as the initial training dataset. For each subsequent round, a given p number of noisy labeled images for each class is selected by the image selection phase to supplement the current training dataset. The best model for each learning round is selected using Equation (1).

3.2.2 Noisy image selection phase

In this second phase, the features that represent the images in the current training set are extracted from the convolutional part of the neural network that is currently being learned. The representation of each class (*i.e.*, class centroid) is then calculated using Equation (3). The features of each noisy image are extracted using the convolutional part of the current neural network to estimate the label of the image using Equation (2). The corresponding score of each remaining noisy labeled image in a particular round is then calculated by Equation (4). The p images for each class are ranked and selected according to their score (*i.e.*, high to low) to add to the current training set for the next learning round.

The key component of this learning framework is incremental enrichment of the training set with the selected noisy images, which may convey additional information useful in improving the generalization of a model being built. As a consequence, the parameters of the neural network, as well as the class centroids, are continuously updated, thereby affecting the estimated labels of noisy images. The intuition is that the information obtained from noisy labeled images should improve the prediction performance of the learned classifier up to a certain point. That is, the hyperparameter p and the number of training rounds both play an important role in controlling the amount of signal and noise in the training set. If they are too low, the classification performance may not be improved, as the amount of new information may not be sufficient. In contrast, the high number of training rounds or the noisy images selected in each round would derail classification performance due to the cumulative amount of noise injected into the training set.

3.3 Multi-round learning algorithm

As shown in Algorithm 1, the training and noisy image selection phases are carried out iteratively. An initial classifier is trained from small ground truth images in the training phase. The image selection phase then proceeds to augment the current training set with selected noisy images whose labels are estimated. Since the feature extractor in the second phase shares the parameters of the resulting network $\mathcal{F}$ learned in the first phase. Thus, the images selected in the first round would have features similar to the ground truths. As the learning process continues, the features extracted in the current round may deviate from the former round as a result of an increasing number of noisy examples in the training set, thereby directly impacting the classification performance.

Algorithm 1 Multi-round learning with noisy labeled data

Input: $\boldsymbol{S}$, $\mathcal{D}$, R, α, p, $\mathcal{F}(\theta_0)$, epochs

$r \leftarrow 1$ ▷ round counter
$T_r \leftarrow \boldsymbol{S}$ ▷ Initial training set
do
  $\mathcal{F}(\theta_r) \leftarrow$ train and validate($\mathcal{F}(\theta_{r-1})$, T_r, epochs)
  $\mathbf{Z} \leftarrow$ compute centroid($\mathcal{F}(\theta_r)$, T_r)
  for each $x_i, y_i \in \mathcal{D}$ do
    $\hat{y}_i \leftarrow$ label correction($\mathcal{F}(\theta_{r-1})$, T_r, $\mathbf{Z}$)
    $\phi_i \leftarrow$ compute score($\mathcal{F}(\theta_r)$, $i, x_i, y_i, \hat{y}_i, \mathbf{Z}, \alpha$)
  end for
  Order images in each class
  $D_r \leftarrow$ Select top p images from each cluster
  $T_{r+1} \leftarrow T_r \cup D_r$
  $\mathcal{D} \leftarrow \mathcal{D} \backslash D_r$
  $r \leftarrow r + 1$
while $r \leq R$

A given deep neural network architecture, $F(\theta_0)$, is learned and then validated through *train and validate*, using the corresponding training set for each round and a validation set, which remains constant throughout the learning process. During this step, the validation error of the learned classifier is monitored, so the best model is selected for later use. A set of features of a given image is extracted by flattening the final layer of the learned convolutional part of the network architecture. The centroids of the classes are repeatedly updated in *compute centroid*. For each training round, the entire set of remaining noisy images is explored to find potential candidates to be added to the current training set. That is, each noisy image is unnecessarily evaluated many times as long as it does not get selected. This exploration method is trivial and becomes time consuming. To reduce such an overhead, we may consider making use of either deliberate or random scheme to select a set of images being evaluated for each round or exercising an early stopping mechanism with respect to a certain condition, such as performance on validation set or centroid variation.

4. EXPERIMENTAL SETUP

This section presents the setup of our experiments aiming to study the behavior of our proposed iterative learning framework when assigning various values to its hyperparameters.

4.1 Ground truth image acquisition

As mentioned above, this work focuses on four classes of diseases and pests: Algal spot, anthracnose, mealybugs, and scale pit, which are common threats to Durian in tropical farming areas where this study was conducted. To construct this ground truth dataset, domain experts were asked to take photos of Durian leaves that were infected by diseases or pests and then assign a class label to each photo taken accordingly. The photos were taken without any restrictions so that they have non-homogeneous backgrounds and leaves' positions, different illumination conditions, various stages of disease development, etc. [49] (see Fig. 1), which should be similar to what the farmers would have. With limited time and resources, there are 400 expert-labeled images (100 images per class), which is not sufficient to yield us a good performance CNN-based classifier. This small ground truth dataset is further divided into training, validation, and test sets used to learn, select, and test the resulting classifier. Let us denote these three cleanly labeled subsets by $\boldsymbol{\mathcal{S}}$, $\boldsymbol{\mathcal{S}}_v$, and $\boldsymbol{\mathcal{S}}_t$, respectively. Note that of 100 expert-labeled images for each class, we have $|\boldsymbol{\mathcal{S}}| = 50$, $|\boldsymbol{\mathcal{S}}_v| = 20$, and $|\boldsymbol{\mathcal{S}}_t| = 30$.

4.2 Collecting noisy images using search engine

In order to enrich a small ground truth training set, we make use of noisy labeled images from online sources to help increase model generalization and overcome the problem of overfitting [42]. The label assigned to an image posted on the Internet is noisy as the result of various factors, especially a person labeling an image may not be an expert, thereby one may confuse the symptom appeared on the leaves from a certain disease or pest with another. In addition to label noise, another major complication posed by using images retrieved via a Web search engine is their quality, as the search results of a given query string may contain many irrelevant images. Even though label noise of online images presents a huge challenge to train deep learning models, the large amount of such images may provide supplement information to the small ground truth, leading to a better representation and then classification.

A set of noisy images for class c denoted by $\mathcal{D}_c$ is collected by constructing a set of corresponding query strings denoted by $\boldsymbol{Q_c}$, in which each query string $\boldsymbol{q} \in \boldsymbol{Q_c}$ is inputted to the Bing search engine via Python library of the Bing image downloader[6] to look for the top 100 images relevant to q. To obtain a sufficient amount of noisy labeled images, for each class, we construct at least five corresponding query strings with various degrees of specificity. For example, the query strings of class "*Anthracnose*" are "*Anthracnose*", "*Durian anthracnose*", "*Leaf anthracnose*", "*Durian leaf anthracnose*", and "*Anthracnose fungal disease in plants*". Thus, $\boldsymbol{\mathcal{D}_c} =_{q \in Q_c} \boldsymbol{I_q}$ contains at least 500 images, where $\boldsymbol{I_q}$ is a set of the top 100 images that are relevant to the input query q obtained from the Bing search engine. Note that all sets of search result images are used as is, without removing redundant or irrelevant images; hence the set of noisy labeled images for all four classes is defined by $\boldsymbol{\mathcal{D}} = U_{c \in C} \boldsymbol{\mathcal{D}_c}$, where $\mathbf{C}$ is a set of four diseases and pests of Durian leaves particular to this study.

4.3 Deep neural network architecture

As mentioned above, our ultimate goal is to develop a mobile application that accurately classifies a photo taken by a farmer to a corresponding class of Durian disease or pest, which should allow farmers to conveniently identify a problem before it is too late without seeking or waiting any help from domain experts. Therefore, in our experiment, we chose to build a classifier from MobileNet, which is a popular CNN neural network architecture for image processing applications used on devices with limited resources. The convolutional part of this architecture is pre-trained on the ImageNet dataset except for several last layers in order to fine-tune their parameters. The classification part consists of two fully connected layers with 0.5 dropout rate; the RMSprop is used for optimization with a fixed learning rate at 0.00001. The training process contains 20 epochs with a batch size of 32. We also keep track of the loss and precision of the learned model performed in the training and validation sets, the best model is selected with respect to the validation loss, and then its performance is evaluated on the test set.

4.4 Hyperparameters being studied

Although several hyperparameters are introduced in our proposed learning framework, there are three of them that directly affect the prediction performance of a classifier that is being built. The first two are the number of learning rounds and the number of noisy labeled images selected from each class in each round, denoted by p. Note that both hyperparameters are used to control the amount of signal and noise to be added to the current training set. In this work, we study only the effect of p at 1, 5, and 10, thus keeping the number of learning rounds constant at 11 (*i.e.*, noisy images are used for only 10 rounds) to also ensure that the learning process will end after reaching the specific round (see Algorithm 1). Another hyperparameter to study here is α, which is

[6] https://pypi.org/project/bing-image-downloader/

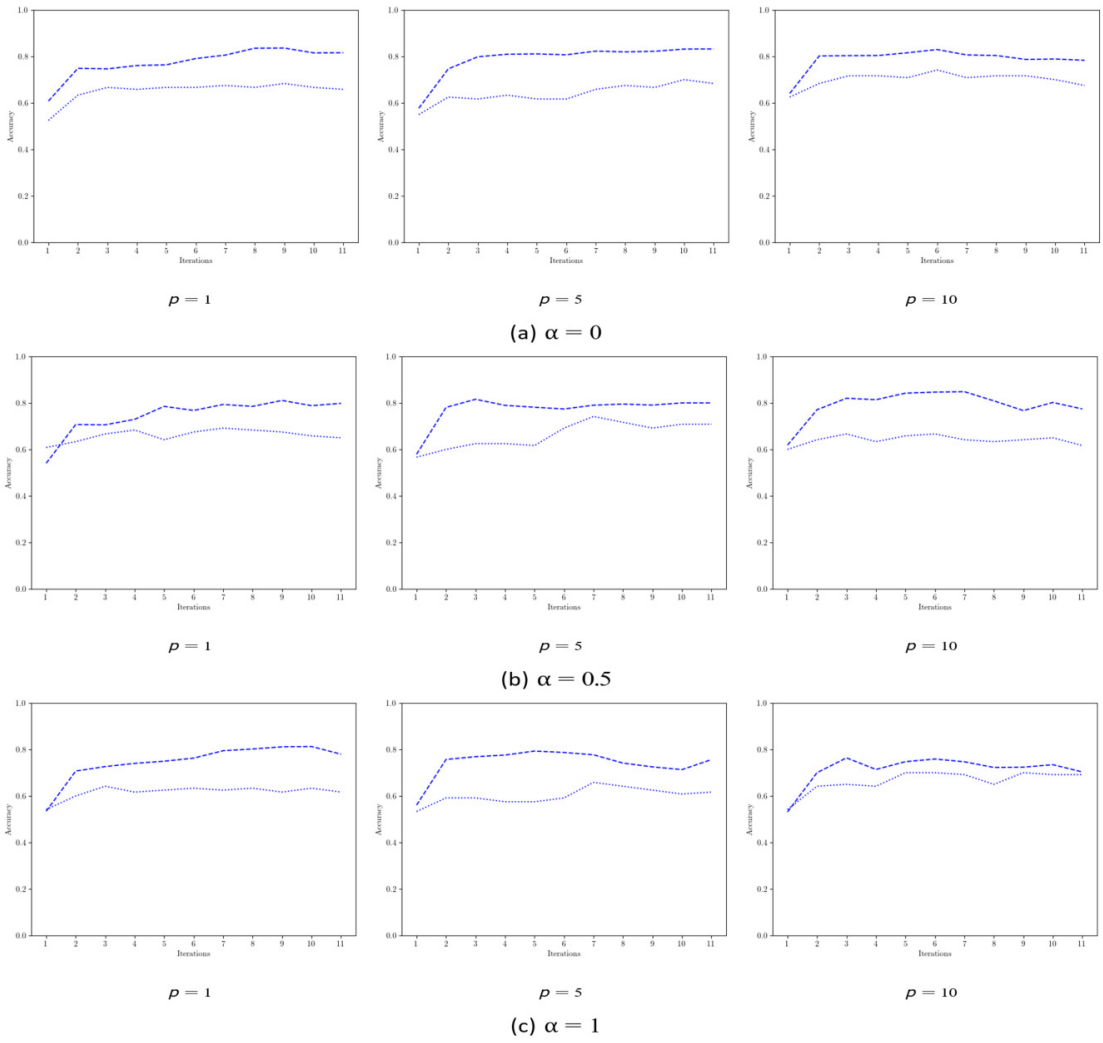

Fig.4: *The plots compare the prediction accuracy of models learned from various assignments of hyperparameters p and on the validation set (dashed lines) and the test set (dotted lines). By leveraging noisy images obtained from search engines, the classifications' performance is improved, which not only indicates that such images contain useful information but also demonstrates the effectiveness of the proposed image selection scheme. The best model is obtained in the seventh iteration of $p = 5$ and $\alpha = 0.5$, as the prediction accuracy of the test set is increased by 20% from the initial model that is learned purely on a small ground-truth training set (i.e., the first iteration).*

used in Equation (4) to compute the score for each remaining noisy label image in each learning round. This hyperparameter is used to determine the trust we have in the label originally assigned to the image. In other words, given a noisy labeled image, if the label estimated by the framework is not identical to the original label assigned by the online source, then the score of the image depends on the trustworthiness of the online source. We carried out our experiments by assigning α to 0, 0.5, and 1. It should be noted that while p is responsible for the quantity of noisy images added to the training set; α reflects the quality of the images through the scoring function.

5. EXPERIMENTAL RESULTS AND DISCUSSION

Various models were built using different combinations of values assigned to the hyperparameters p and α. The prediction accuracy of the classifier obtained from each learning round was assessed with validation and test sets. The evaluation results are presented in Fig. 4.

We can see the general behavior of the classifiers built in various settings of p and α, in which the initial model always has the worst prediction accuracy in the validation set, as it was only trained from the small ground truths. As the training goes on, the number of noisy labeled images in the training set keep increasing, the prediction accuracy of the classifiers on the

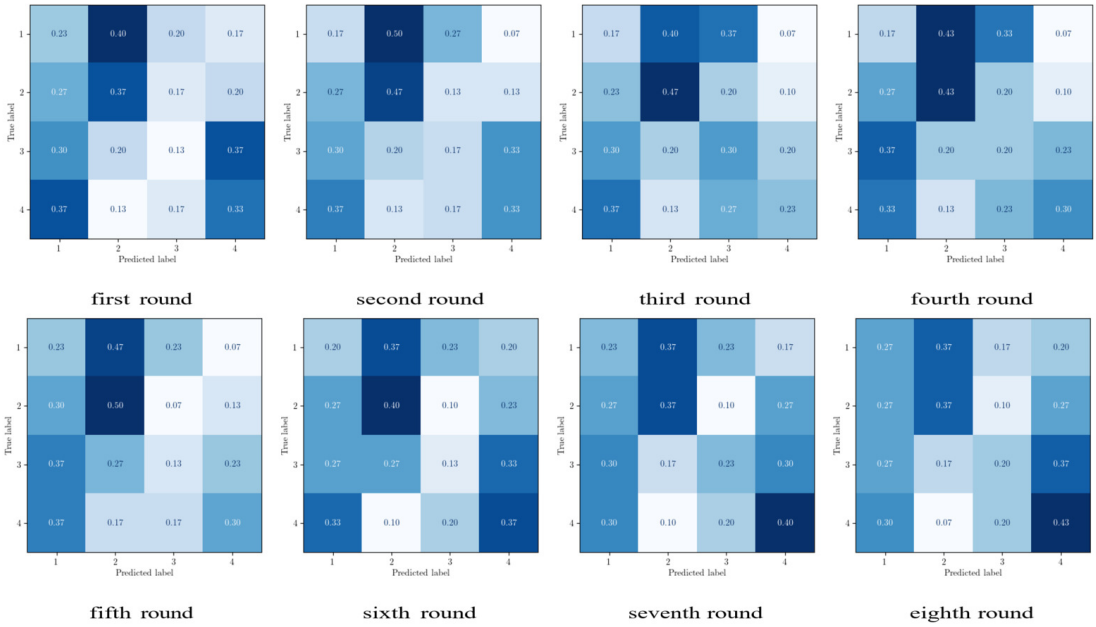

Fig.5: *The confusion matrices illustrate the prediction accuracy of the classifier learned from the best combination of the hyperparameters on particular classes for eight learning rounds (out of 11). Note that the second and fourth classes are easier to classify, as the accuracy on these classes is always higher than that of the other two classes. Although the prediction accuracy for each class is just fair, adding noisy images to the training set provides additional information that can help the learned classifier generally become less confused.*

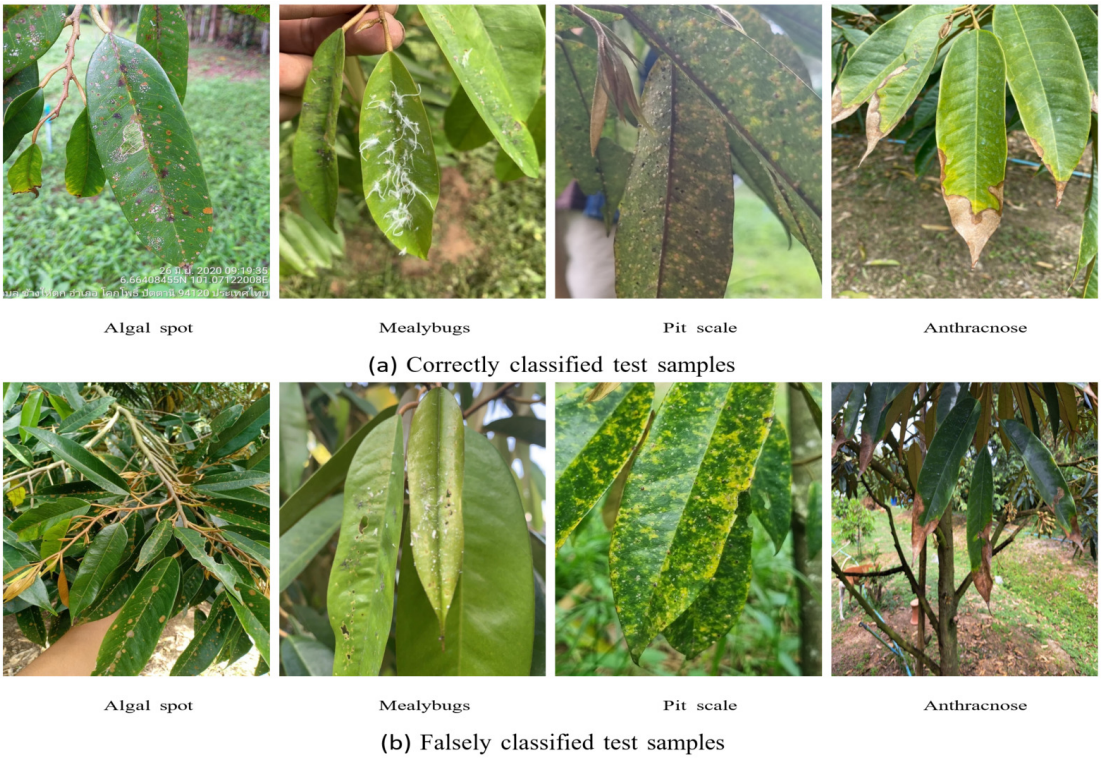

Fig.6: *The test samples used to evaluate the best-performing classifier are shown here. The images in the top row were correctly classified into their respective classes, while those in the bottom row were misclassified. Notably, some of the misclassified images are so ambiguous that even humans would find it difficult to assign them to the correct class.*

test set increases typically after the fourth learning round and then gradually decreases after the eighth round. This behavior indicates that noisy images provide useful information to improve the generalization of a model being built. However, when the number of noisy labeled images in the training set increases to a certain point, until there is no new information to be conveyed, and the amount of noise incurred pollutes and overwhelms the training set, thereby lowering the overall performance of the classifiers.

When we consider each hyperparameter separately, we can see that adding one noisy image in each learning round ($p = 1$) results in a poor performance classifier, as the amount of useful information from the added imageis not sufficient to improve the generalization of the model. Consequently, more learning rounds may be required to produce a high performance model, which is unnecessarily expensive to train. In contrast, if a large number of noisy labeled images is added to the training set in each learning round ($p = 10$), there is a greater chance that a lot of noise accumulates and corrupts the training set in the first few rounds, making the search for a good classifier impossible. By adding only 5 noisy labeled images to the training set in each learning round ($p = 5$), we generally obtain better classifiers than in the former cases, as the number of noisy images being added is large enough to provide a significant amount of useful information, but not too large to accumulate an unacceptable amount of noise.

When examining only α, which is the important hyperparameter to calculate the score for each noisy labeled image, good quality images are selected to augment the training set in each round. Note that the higher the value assigned to α, the greater confidence is placed in the original label; otherwise, the estimated label is more preferable. That is, in the first case in which $\alpha = 0$, only the estimated label is taken into account to compute the score if it is not in accordance with the original label. On the other hand, $\alpha = 1$ means that only the original label is considered when calculating the score of each noisy image if it does not conform to the estimated label.

When comparing the prediction performance of the classifiers produced from these two extreme cases, we can see that the former case typically yields the better model than the latter, indicating that the label estimated by using features extracted from the learned convolutional part of the deep neural network is more accurate than the original labels retrieved from the web that were used in training. However, the best classifier was built by equally weighing the label information from both sources ($\alpha = 0.5$). This result demonstrates that the original labels are not purely noise but rather require validation from another source of information. By incorporating votes from both sources, the model assigns a better score to noisy images. It should be noted that if we know the retrieved noisy images come from trusted online sources, a higher value of α could be used to reflect the trustworthiness of these image sources. We further analyze the prediction performance of our best classifier for each class, as shown in Fig. 5. The round-by-round confusion matrix, obtained from testing the model trained with the optimal hyperparameters ($p = 5$ and $\alpha = 0.5$) on the test set, reveals that the second and fourth classes are easier to classify compared to the other two. This is because images of these diseases and pests typically exhibit clear features. By carefully selecting noisy images from the internet and adding them to the training set in an iterative manner, the classifier's confusion for each class is reduced to some extent, indicating that these images contain valuable information to improve model generalization. However, a significant gap in accuracy remains, as many of the online images are of poor quality (see Fig. 2).

From the confusion matrices in Fig. 5, we further examine the test samples used to evaluate our learned classifier. We selected two test examples for each class: one correctly classified and one misclassified. The selected images for all classes are presented in Fig. 6. The images in the top row show the test samples that our model correctly classified, while the misclassified images are displayed in the bottom row. As we can see, the correctly classified images of each class have clear features of the corresponding diseases and pests on the Durian leaves. In contrast, the misclassified images often contain unclear or ambiguous features, which caused the classifier to make incorrect predictions. For example, the second and third images were mistakenly classified as *Algal Spot*, since the features of *Mealybugs* and *Pit Scale* are not clearly visible on the leaves. Additionally, confusion can arise when the point of interest in the image is uncertain. For instance, the first image was wrongly assigned to *Anthracnose*, while the fourth image was labeled as *Pit Scale*. This issue can be easily mitigated by advising farmers using our mobile application to ensure that the photo focuses on the clear symptoms of diseases or pests on the Durian leaves.

6. CONCLUSION

This work presents the multi-round learning framework that makes use of noisy labeled images obtained from the Internet as an additional source of information to improve the generalization of a model being built. The bottom line is that the proposed framework is sufficiently general for building classifiers from any CNN architecture to support various classification tasks. Enriching the training set with a suitable number of noisy images in each learning round to introduce new and useful information with a small amount of noise incurred, the general prediction accuracy of the classifier can be improved compared to the initial model, which is purely learned from

small ground-truth examples. Although the photos retrieved from the Internet may be redundant and of poor quality, the evaluation results have shown that our approach is promising, especially in the situation where there is a limited number of expert-labeled images. As the scarcity of high-quality labeled images occurs largely in many domains, weakly supervised techniques as described in this work will become an important approach to address this challenge and enable us to build high-performance models for various tasks.

For future work, our aim is to extend the scope of this work by including not only other leaf diseases and pests but also those whose symptoms appeared in the other parts of Durian. We will apply this framework to other local crops that are as economically important as Durian. A mobile application must be developed using a model built from the proposed learning framework so that its effectiveness can be practically assessed through prediction accuracy and user satisfaction.

Improving the proposed framework is another future goal, especially accelerating the image selection phase in the training process. Also, it is worth noting that the resulting model is not optimal; this issue may be addressed by combining information from both phases of the framework into one optimization function, which must be efficiently solved for an optimal solution. Moreover, we may use the adversarial generative approach by combining information from collected ground truths and noisy images to generate additional samples for classes in which the resulting classifier performs poorly.

ACKNOWLEDGEMENT

The authors acknowledge the financial support of this work from the National Science and Technology Development Agency (NSTDA) Coordinating Center for Thai Government Science and Technology Scholarship Students (CSTS). The ID of this funding is JRA-C0-2565-16959-TH.

AUTHOR CONTRIBUTIONS

Conceptualization, S.J.; methodology, S.J.; software, S.J. and K.R.; validation, S.J.;data curation, S.J.; writing—original draft preparation, S.J. and K.R.; writing—review and editing, S.J. and K.R.; visualization, S.J.; funding acquisition, S.J. All authors have read and agreed to the published version of the manuscript.

References

[1] S. D. Khirade and A. B. Patil, "Plant Disease Detection Using Image Processing," *2015 International Conference on Computing Communication Control and Automation*, Pune, India, pp. 768-771, 2015.

[2] M. Dutot, I. Nelson and R. Tyson, "Predicting the Spread of Postharvest Disease in Stored Fruit, with Application to Apples," *Postharvest Biology and Technology*, vol. 85, pp. 45–56, Nov. 2013.

[3] Sabarre, A. Navidad, D. Torbela and J. Adtoon, "Development of Durian Leaf Disease Detection on Android Device," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, pp. 4962–4971, 2021.

[4] J. Al Gallenero and J. Villaverde, "Identification of Durian Leaf Disease Using Convolutional Neural Network," *2023 15th International Conference on Computer and Automation Engineering (ICCAE)*, Sydney, Australia, pp. 172-177, 2023.

[5] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 779-788, 2016.

[6] L. Pang, Y. Lan, J. Guo, J. Xu, and X. Cheng, "Deeprank: A New Deep Architecture for Relevance Ranking in Information Retrieval," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 257–266, 2017.

[7] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 328–339, 2018.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT 2019*, pp. 4171–4186, 2019.

[9] W. Jing and C. S. Lim, "Multi-granularity Self-attention Mechanisms for Few-shot Learning," *ECTI-CIT Transactions*, vol. 18, no. 4, pp. 522–530, 2024.

[10] Dey, S. Biswas and L. Abualigah, "Efficient Violence Recognition in Video Streams using ResDLCNN-GRU Attention Network," *ECTI-CIT Transactions*, vol. 18, no. 3, pp. 329–341, 2024.

[11] Dey and S. Biswas, "Shot-vit: Cricket Batting Shots Classification with Vision Transformer Network," *International Journal of Engineering*, vol. 37, no. 12, pp. 2463–2472, 2024.

[12] Calvo, S. Calderon-Ramirez, J. Torrents-Barrena, E. Munoz and D. Puig, "Assessing the Impact of a Preprocessing Stage on Deep Learning Architectures for Breast Tumor Multi-class Classification with Histopathological Images," in *Latin American High Performance Computing Conf.*, Springer, pp. 262–275, 2019.

[13] V. Iglovikov, A. Rakhlin, A. Kalinin and A.

Shvets, "Paediatric Bone Age Assessment Using Deep Convolutional Neural Networks," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, pp. 300–308, 2018.

[14] X. Yu, X. Wu, C. Luo and P. Ren, "Deep Learning in Remote Sensing Scene Classification: A Data Augmentation Enhanced Convolutional Neural Network Framework," *GIScience and Remote Sensing*, vol. 54, no. 5, pp. 741–758, 2017.

[15] Tan, F. Sun, T. Kong, W. Zhang, C. Yang and C. Liu, "A Survey on Deep Transfer Learning," in *International Conference on Artificial Neural Networks*, Springer, pp. 270–279, 2018.

[16] T. Xiao, T. Xia, Y. Yang, C. Huang and X. Wang, "Learning from massive noisy labeled data for image classification," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, pp. 2691-2699, 2015.

[17] Olivier Chapelle; Bernhard Schölkopf; Alexander Zien, "An Augmented PAC Model for Semi-Supervised Learning," in *Semi-Supervised Learning* , MIT Press, 2006, pp.397-419.

[18] S. Calderon-Ramirez, S. Yang and D. Elizondo, "Semisupervised Deep Learning for Image Classification with Distribution Mismatch: A Survey," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 6, pp. 1015–1029, 2022.

[19] H. Song, M. Kim, D. Park, Y. Shin and J. -G. Lee, "Learning From Noisy Labels With Deep Neural Networks: A Survey," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8135-8153, Nov. 2023.

[20] K. -H. Lee, X. He, L. Zhang and L. Yang, "CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 5447-5456,2018.

[21] H. Song, M. Kim and J. Lee, "SELFIE: Refurbishing Unclean Samples for Robust Deep Learning," in *Proceedings of the 36th International Conference on Machine Learning*, pp. 5907–5915, 2019.

[22] Nguyen, C. Mummadi, T. Ngo, T. Nguyen, L. Beggel and T. Brox, "SELF: Learning to Filter Noisy Labels with Self-ensembling," in *International Conference on Learning Representations*, 2020.

[23] Nettleton, A. Orriols-Puig and A. Fornells, "A Study of the Effect of Different Types of Noise on the Precision of Supervised Learning Techniques," *Artificial Intelligence Review*, vol. 33, no. 4, pp. 275–306, 2010.

[24] G. Patrini, A. Rozza, A. K. Menon, R. Nock and L. Qu, "Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 2233-2241, 2017.

[25] Hendrycks, M. Mazeika, D. Wilson and K. Gimpel, "Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 10477–10486, 2018.

[26] R. Wang, T. Liu and D. Tao, "Multiclass Learning with Partially Corrupted Labels," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2568–2580, 2017.

[27] H. Chang, E. Learned-Miller and A. McCallum, "Active Bias: Training More Accurate Neural Networks by Emphasizing High Variance Samples," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1003–1013, 2017.

[28] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan and A. Rabinovich, "Training Deep Neural Networks on Noisy Labels with Bootstrapping," in *International Conference on Learning Representations*, 2015.

[29] X. Ma, Y. Wang, M. Houle, S. Zhou, S. Erfani, S. Xia, S. Wijewickrema and J. Bailey, "Dimensionality-driven Learning with Noisy Labels," in Proc. ICML, 2018.

[30] P. Chen, J. Ye, G. Chen, J. Zhao and P. Heng, "Beyond Class-conditional Assumption: A Primary Attempt to Combat Instance-dependent Label Noise," in *The Thirty-Fifth AAAI Conference on Artificial Intelligence*, pp. 11442-11450, 2021.

[31] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu and D. Meng, "Meta-Weight-Net: Learning an Explicit Mapping for Sample Weighting," in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, pp. 1917–1928, 2019.

[32] Z. Wang, G. Hu and Q. Hu, "Training Noise-Robust Deep Neural Networks via Meta-Learning," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 4523-4532, 2020.

[33] E. Malach and S. Shalev-Shwartz, "Decoupling When to Update from How to Update," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, pp. 960–970, 2017.

[34] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang and M. Sugiyama, "Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels," in *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada, pp. 8527–8537, 2018.

[35] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang and M. Sugiyama, "How Does Disagreement Help Generalization Against Label Corruption?," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, California, USA, 2019.

[36] P. Chen, B. Liao, G. Chen and S. Zhang, "Understanding and Utilizing Deep Neural Networks Trained with Noisy Labels," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, California, USA, 2019.

[37] J. Huang, L. Qu, R. Jia and B. Zhao, "O2U-Net: A Simple Noisy Label Detection Approach for Deep Neural Networks," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 3325-3333, 2019.

[38] J. Han, P. Luo and X. Wang, "Deep Self-Learning From Noisy Labels," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 5137-5146, 2019.

[39] J. Li, R. Socher and S. Hoi, "DivideMix: Learning with Noisy Labels as Semi-supervised Learning," in *International Conference on Learning Representations*, 2020.

[40] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver and C. Raffel, "MixMatch: A Holistic Approach to Semi-supervised Learning," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 5050–5056, 2019.

[41] M. Pathan, N. Patel, H. Yagnik and M. Shah, "Artificial Cognition for Applications in Smart Agriculture: A Comprehensive Review," *Artificial Intelligence in Agriculture*, vol. 4, pp. 81–95, 2020.

[42] R. Abbasi, P. Martinez and R. Ahmad, "Crop Diagnostic System: A Robust Disease Detection and Management System for Leafy Green Crops Grown in an Aquaponics Facility," *Artificial Intelligence in Agriculture*, vol. 10, pp. 1–12, 2023.

[43] P. Bedi and P. Gole, "Plant Disease Detection Using Hybrid Model Based on Convolutional Autoencoder and Convolutional Neural Network," *Artificial Intelligence in Agriculture*, vol. 5, pp. 90–101, 2021.

[44] A. Paymode and V. Malode, "Transfer Learning for Multi-crop Leaf Disease Image Classification Using Convolutional Neural Network VGG Neural Network," Artificial Intelligence in Agriculture, vol. 6, pp. 23–33, 2022.

[45] A. Abbas, S. Jain, M. Gour, and S. Vankudothu, "Tomato Plant Disease Detection Using Transfer Learning with C-GAN Synthetic Images," *Computers and Electronics in Agriculture*, vol. 187, no. 106279, 2021.

[46] C. Liu, H. Zhu, W. Guo, X. Han, C. Chen, and H. Wu, "EFDet: An Efficient Detection Method for Cucumber Disease under Natural Complex Environments," *Computers and Electronics in Agriculture*, vol. 189, no. 106378, 2021.

[47] M. Mathew and T. Mahesh, "Leaf-based Disease Detection in Bell Pepper Plant Using YOLO v5," *Signal, Image and Video Processing* , vol. 16, pp. 841–847, 2022.

[48] A. 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.*

[49] K. Jha, A. Doshi, P. Patel and M. Shah, "A Comprehensive Review on Automation in Agriculture Using Artificial Intelligence," *Artificial Intelligence in Agriculture*, vol. 2, pp. 1–12, 2019.

Sasin Janpuangtong is an instructor with the Department of Computer Engineering at ChiangMai University, Chiang Mai, Thailand. His research interests include intersection of AI and cybersecurity and explainable AI, with main contributions in weak supervisions, threat intelligence, and malware analysis. Janpuangtong received a Ph.D. in computer science from Texas A&M University. He is a director of cybersecurity track at Chiang Mai University. Contact him at sasin.ja@cmu.ac.th.

Kritwara Rattanaopas is an instructor in the Department of Computer Engineering at Prince of Songkla University, Songkhla, Thailand. His research interests include the design and development of chatbots, big data, data engineering, natural language processing, large model languages, and knowledge graphs (ontology). His main contributions are in high-performance computing and cloud computing. Dr. Rattanaopas received a Ph.D. in Computer Engineering from Prince of Songkla University.