SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage: https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Non-Parametric Machine Learning for Pollinator Image Classification: A Comparative Study

Nurul Shuhada Zamri[1], Nurul Amelina Nasharuddin[1,*]

[1] Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 11 July 2023
Received in revised form 12 October 2023
Accepted 1 November 2023
Available online 26 November 2023

Keywords:
Pollinators; Image Classification; Non-Parametric Machine Learning; Shape and Colour Features; Random Forest Classifier

ABSTRACT

Pollinators play a crucial role in maintaining the health of our planet's ecosystems by aiding in plant reproduction. However, identifying and differentiating between different types of pollinators can be a difficult task, especially when they have similar appearances. This difficulty in identification can cause significant problems for conservation efforts, as effective conservation requires knowledge of the specific pollinator species present in an ecosystem. Thus, the aim of this study is to identify the most effective methods, features, and classifiers for developing a reliable pollinator classifier. Specifically, this initial study uses two primary features to differentiate between the pollinator types: shape and colour. To develop the pollinator classifiers, a dataset of 186 images of black ants, ladybirds, and yellow jacket wasps was collected. The dataset was then divided into training and testing sets, and four different non-parametric classifiers were used to train the extracted features. The classifiers used were the k-Nearest Neighbour, Decision Tree, Random Forest, and Support Vector Machine classifiers. The results showed that the Random Forest classifier was the most accurate, with a maximum accuracy of 92.11% when the dataset was partitioned into 80% training and 20% testing sets. By developing a reliable pollinator classifier, researchers and conservationists can better understand the roles of different pollinator species in maintaining ecosystem health. This understanding can lead to better conservation strategies to protect these important creatures, ultimately helping to preserve our planet's biodiversity.

1. Introduction

Pollinators (mainly insects) are one of the most important things in this world. Humans and all terrestrial ecosystems would collapse if pollinators were not present. Almost 80 percent of the 1,400 crop plants farmed around the world, i.e., those that provide all of our food and plant-based industrial products, require animal pollination as stated in [1,2]. Pollinators, as a result, play an important role in regulating ecosystem services that support food production, habitat, and natural

* *Corresponding author.*
E-mail address: nurulamelina@upm.edu.my

https://doi.org/10.37934/araset.34.1.106115

resources, according to several authors [2-4]. Populations of both wild and managed pollinators were also under threat in many parts of the world thus impacted the human food security reported by [5].

One of the most pressing issues confronting humanity today is the numbers of insects are declining in abundance and diversity, but their population trends remain uncertain as insects are difficult to monitor as stated by Bjerge, Mann and Høye [6]. The trapping and subsequent species identification processes in manual methods take a significant amount of time. According to Fereira *et al.,* [7], taxonomic identification of pollinators especially the bees which depends on microscopic characteristics are a challenge since insect taxonomists are scarce nowadays.

Furthermore, a system that can assist individuals, particularly the younger generation, in identifying the types of insects and pollinators is restricted. Even though the agriculture industry has advanced with modern technology, there is still a lack of a system that can assist the young generation in automatically detecting the type of pollinators. Also, most of the existing available systems are intended to classify pests and insects rather than pollinators.

Therefore, to overcome the limitations of traditional taxonomy, an automatic classification of pollinator types based on machine learning is hoped to help people easily distinguished the types of insects and prescribed appropriate contingency plan to slow down the reduction of the insect and pollinator populations. This work is proposed to be the baseline for further identification of pollinators and subsequently more types and species of pollinators can be automatically identified.

This study will identify the most effective methods, features, and classifiers for developing a reliable pollinator classifier for three common types of pollinators that are easily found, namely black ants (*semut hitam*), ladybirds (*kumbang kura-kura*), and yellow jacket wasps (*tebuan kuning*). See Figure 1. All these pollinators are being used in this study to investigate what characteristics they can contribute to defining their type. Essentially, they are being examined primarily on their body traits because the body is the most visible portion of pollinators and has a variety of characteristics such as shape, colour, texture, and many others.

Fig. 1. Black ant (left), ladybird (middle) and yellow jacket wasps (right)

Specifically, this initial study uses two primary features to differentiate between the pollinator types which are shape and colour. Some image processing and pattern recognition techniques were applied, beginning with receiving an input image and continuing through pre-processing, classifying, and producing classification results. Before identifying the input pollinator image according to its type, the image must go through certain pre-processing operations such as resizing and grey scaling.

Following that, the feature extraction process's relevant descriptors will perform their work. To classify the pollinator, the classifiers such as Support Vector Machine (SVM), *k*-Nearest Neighbour (*k*-NN), Random Forest, and Decision Tree were used. The type of proposed pollinator and the accuracy percentage of the classification result is given at the end of this study.

According to article by [1], a pollinator is anything that assists in the transport of pollen from the male part of the flower (stamen) to the female reproductive organ of the same or a different flower (stigma). Certain plants pollinate themselves, while some are pollinated by pollen delivered by water or wind. Other flowers, however, are pollinated by animals and insects such as bees, wasps, moths, butterflies, birds, flies, and tiny mammals such as bats.

Each of the pollinators has its characteristics. A black ant's body length, for example, ranges from 3.4 mm to 5 mm for workers and 15 mm for the queen. Furthermore, black ants have only one waist segment, and their colour is dark brown-black for workers and mid-brown for queens. Next, for the yellow jacket wasp, it has yellow and black stripes zigzagging along its entire body, from head to tummy. Yellow jackets, unlike bees, have smooth, extended bodies rather than fuzzy, squat appearances. A yellow jacket worker is around 0.5 inches long, while a queen might be 0.75 inches long. For ladybirds, the range in size is from 0.3 to 0.4 inches. While they are born black, their mature colour can vary from yellow to red, and the spots on their half-sphere-shaped forewings (wing covers) can also vary. Nowadays, many industries are designing more powerful machine learning algorithms that can process larger and more complex data sets while producing quicker, more effective results on massive scales.

According to the authors Wäldchen and Mäder [8], machine learning is a form of artificial intelligence that can address problems without even being specifically coded to do so. It is also stated that machine learning is specifically useful for extracting information from vast volumes of rapidly increasing data, and it is especially useful for applications in which the data is hard to analyse analytically, such as processing image and video content.

Image classification is one of supervised machine learning, where the process by which a computer analyses an image and determines which 'class' the image belongs to. Image classification is important in agriculture because it is used for a variety of purposes such as environmental change, agriculture, land use or land planning, urban planning, monitoring, spatial mapping, disaster management, and object detection as mentioned by authors in [9-12]. For example, image classification can be used to classify insect species as reported by Tuda and Luna-Maldonado [13] which help the farmer to easily identify what species of insects destroy their farm. Authors in [14] reported a real-time agriculture picture categorisation framework that employed IoT cameras, sensors and mobile applications. Authors in [15] suggested a framework for recognising durian species since the current techniques struggle to distinguish between these species as their skin colour. An automated system for content-based recognition using multiple attributes would be useful for accurately representing and identifying durians.

Based on the research done by researchers in [13], automatic recognition systems developed by digitising taxonomic features and building a classifier model using machine learning, are promising methods for classifying and naming plants and animals. The most popular use of image-based machine learning is species classification. Pest beetles and (*Callosobruchus Chinensis*) and their parasitoids (parasitic wasps; *Anisopteromalus* and *Heterospilus*) are examples of species that have been classified.

Non-parametric machine learning consists of algorithms that do not make any firm assumptions regarding the form of the mapping function [16]. By not making assumptions, they are free to learn any functional form from the training data. The algorithms can become more and more complex with an increasing amount of data. Three examples of the advanced non-parametric machine learning models are *k*-NN, SVM, and random forest [17]. Authors in [18] recommended that the handcrafted simple image processing algorithm should be tried first, due to its simplicity, for the image classification before resorting to advanced and complex versatile machine learning modelling approaches.

According to a research experiment done by researchers [19] that implemented SVM, traditional ways of insect identification are time-consuming, so they used machine learning that uses computer vision techniques and various image segmentation approaches. In this research, they used colour and shape feature extraction. They generated the colour image of the insects by multiplying the binary image element by element by the original image. For the shape, the author detects the size of the pests. The classifier used to detect the pest is SVM. This research used 100 images as a data set, where 20% were used to test and 80% were used to train. The system can achieve an error of less than 2.5% while doing the classification. It is also stated that even minor variations in the size or colour of the target or parasite pest will increase the classification error. That means the size and colour of the pest must be extracted exactly as in the original images.

Authors in [20] had proposed a vision-based counting and recognition system for flying insects in intelligent agriculture. The system uses shape, texture, colour, and Histogram of Oriented Gradients (HOG) feature extraction. For the shape extraction, due to the wide variation in area and diameter of the same species of flying insect, four geometrical characteristics, including sophistication, duty time, eccentricity, and extension rate, are chosen. Grey level L is set to 64 to extract texture features, and scan angle is set to 0, 45, 90, and 135 to extract texture features.

The Grey-Level Co-occurrence Matrix (GLCM) of the R, G, and B colour components is computed. In this research, the image, cell, and block resolutions are set to 64 x 128, 32 x 32, and 64 x 64, respectively, to extract the HOG features. SVM classifier is also used in this research to classify the types of flying insects. 1000 images are used as the dataset in this work. This system can reach a relatively high detection rate and efficiency. Experimental results show that the average classifying accuracy is 90.18%. This work needs many training samples; however, it is difficult to obtain enough samples of certain specific insects during a certain season.

Authors in [21] proposed a vision-based perception and classification of mosquitoes using a support vector machine. Feature extraction that is used in this system is the length of the body and leg and the colour of the mosquito. The trunk width to leg length ratios of segmented regions was calculated using mathematical formulas and distance measurement algorithms. For colour extraction, the colour histogram for each image is extracted. They used three types of SVM and corresponding kernel functions for the classification process. The research which used 400 images of mosquitos, showed a maximum recall of 98% in identifying the insect. However, the author stated that for future research they should provide additional features to boost the classifier's performance and it is also stated that SVM using a linear kernel function is ineffective for classifying mosquitos and other pests.

2. Methodology

2.1 Dataset

In this stage, the collection of the target pollinators images was carried out. The dataset consists of 62 photos for each pollinator type collected from the Internet since there is no suitable dataset for pollinators available, to the best of our knowledge. The small number were taken because the quality of most of the other pictures is not suitable for datasets. As a total 186 photos were acquired as part of the data set preparation.

The types of pollinators that were included in this initial study are black ant, and yellow jacket wasp because they were common species in Malaysia and good samples of images can easily be found. Furthermore, these three pollinators can be distinguished by features such as colour and shape. This phase also includes improving the image's quality, such as cropping the image and removing the background image before it is accepted as the input image. See Figure 2.

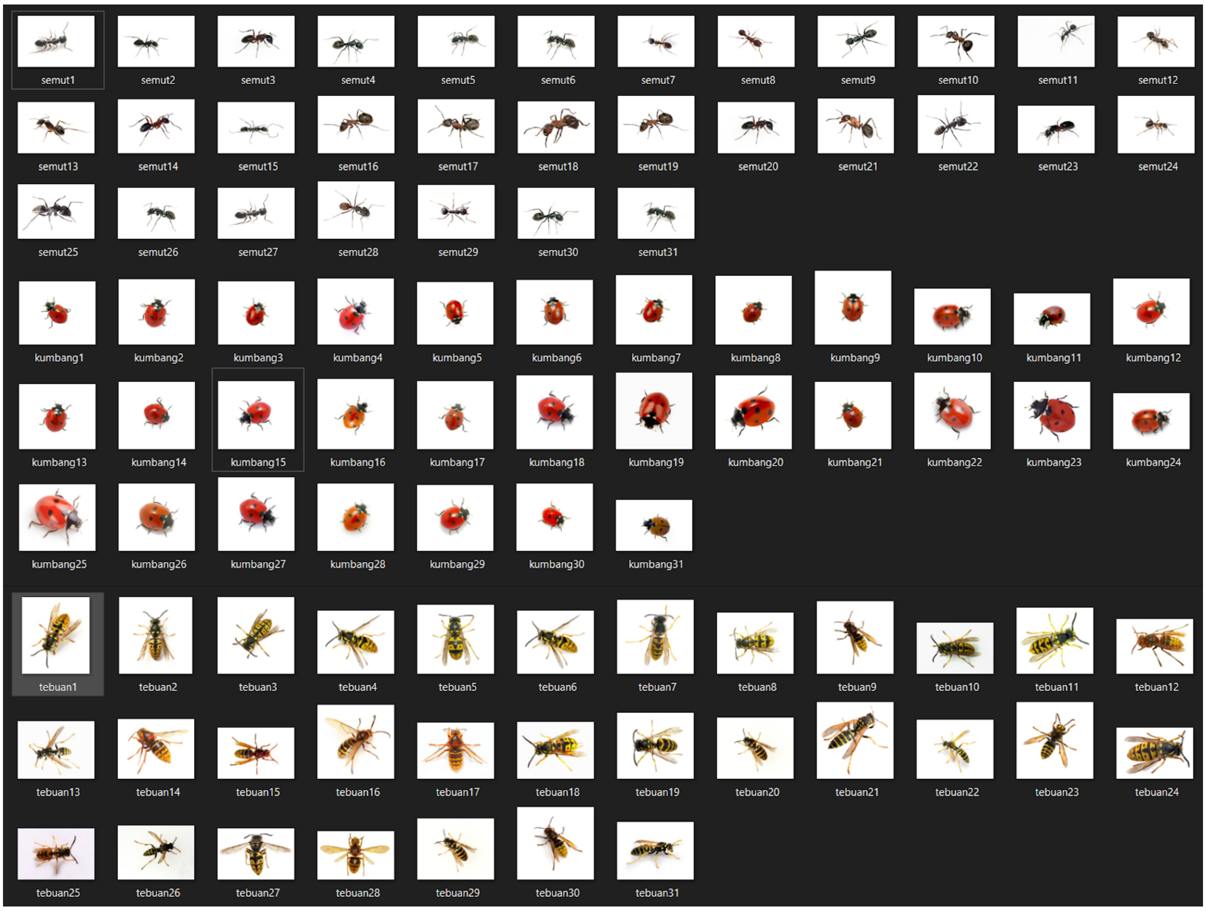

Fig. 2. Samples of the training and testing dataset

2.2 Feature Extraction

The classification begins with the pre-processed of the dataset's images with each image was resized into size 150 x 150 and converted into a grayscale image using the image thresholding. The greyscale images were then binarized by calculating the threshold using Otsu's algorithm. Values of the pixels that are greater than the threshold are replaced with the 1 (white) and other pixels are replaced with 0 (black). The binarized image is smoothed by filtering noises using a rectangular averaging filter of size 3 × 3. Then, the image was convolved using a 3 × 3 Laplacian operators to get the margin of the image.

The processed images then went through the shape feature extraction procedure where the area of the pollinators was taken. The input images were also gone through the morphological process to remove any edge touching that may interfere with shape extraction. This pre-processing step is critical because the classification's outcome is dependent on this phase. Shapes and colours were chosen as features to extract since the differences in shape and colour between these three types of pollinators are the most obvious. To extract such features from the images, proper descriptors must be used so that the classification result is as predicted.

2.2.1 Shape feature extraction

This study analysed labelled regions in images and extract shape-related features for image classification tasks, which in this case the area of the labelled regions. Pollinator's image area gives the total number of pixels for each smoothed pollinator and every pollinator has its own characteristics. This feature was combined with the colour feature and was used as the input to the machine learning model.

2.2.2 Colour feature extraction

Colour is being chosen as another feature in addition to shape because some the pollinators have different colours. The colour histogram feature was extracted from the pollinator's images. The distribution of colours in an image is considered by the colour histogram. The approach has the advantage of not being influenced by any rotation or translation of the image. A histogram provides a statistical graph that shows the number of pixels of the red, green and blue colours of the image. Based on the Eq. (1) and Eq. (2) below, the feature vector was formed using the mean (μ) and standard deviation (σ) of the intensity values of the red, green, and blue channels:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

$$\sigma = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right)} \tag{2}$$

where *N* is the total number of pixels within an image and x_i is the current pixel being processed. Figure 3 shows the process flow of the proposed multi-feature-based extraction.

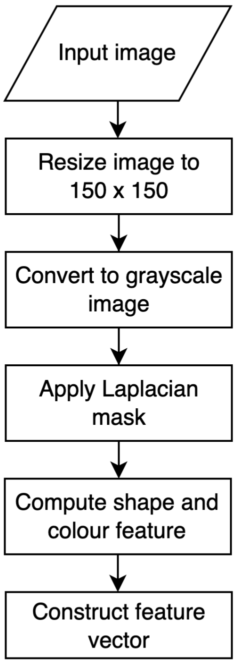

Fig. 3. Process flow for shape and colour feature extraction

2.3 Training and Testing

After undergoing the feature extraction process, the images were divided into training and testing processes. Figure 4 depicts the flow chart for the training and testing process for the classification. The training process took about 80% of the images in the dataset, while for the testing 20% of the images were taken. For the training process, classifiers such as SVM, KNN, Random Forest, and Decision Tree were used. Each of the classifiers will classify the image, and the classifier with the highest accuracy value is chosen as the classifier for testing. The testing images undergo the classification process, and the result of the classification is displayed along with the accuracy of the classification. The source codes for the classification algorithms in Google Colab using the Python language. The algorithm for classification is built by applying image processing and pattern recognition techniques.

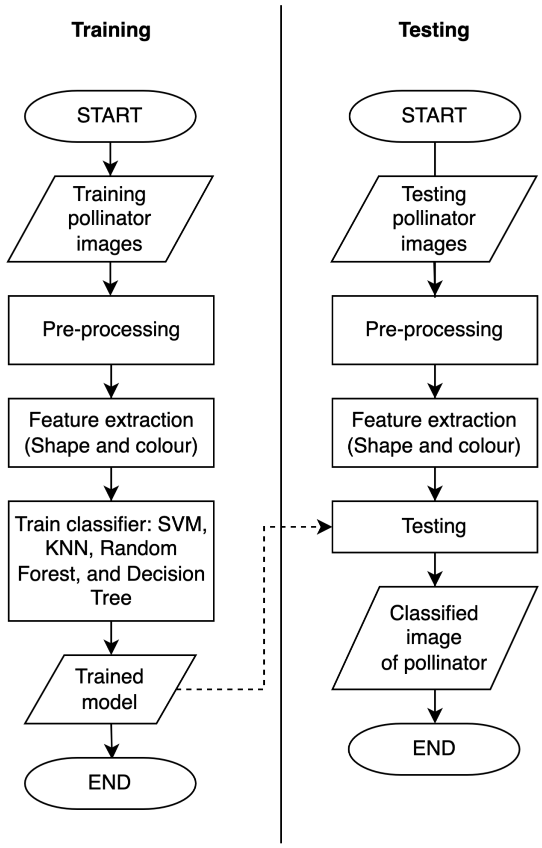

Fig. 4. Flowchart for the training and testing

3. Results and Discussion

After the dataset preparation and training phases, the algorithms were being tested using the best non-parametric classifier. The results were being analysed in detail and discussed.

3.1 Accuracy Comparison of the Features

For the classification, two features were chosen, namely the shape and colour of the pollinators. The area was chosen for the shape feature and the width of the pollinator body was measured by its area. Red, green, and blue colour ranges were derived from the pollinator images. SVM, KNN, Random Forest, and Decision Tree classifiers were used in this evaluation. Table 1 displays the results of the comparison between each single feature of colour and shape and combined features.

Table 1
Classifiers accuracy for single and combined features

Features	SVM	KNN	Random Forest	Decision Tree
Colour	68.42%	78.65%	76.32%	60.53%
Shape	55.26%	60.53%	57.89%	55.26%
Colour + shape	73.68%	81.58%	**89.47%**	63.16%

The results above show that the Random Forest classifier gave the best accuracy which is 89.47% for the combination of both features to classify the pollinator images. Another classifier that is considered a good classifier is the KNN as the classifier gives more than 80% accuracy which is 81.58%. SVM and Decision Tree classifiers both respectively give 73.68% and 63.16% of accuracy when done in the classification process.

3.2 Accuracy for Each Type of Pollinators

Table 2 has detailed information on the percentage of accuracy for each type of pollinator using the Random Forest classifier using both colour and shape features. The black ant has the highest classification accuracy of 90.52%, followed by the ladybird with an accuracy value of 87.96%. The last one is the yellow jacket wasp, which has a classification accuracy of 85.61%. The average accuracy of this categorisation is 88.03%. In conclusion, the Random Forest classifier appropriately classifies all three categories of pollinators.

Table 2
Percentage of accuracy for each pollinators type

Pollinator	Accuracy
Black ant	90.52%
Ladybird	87.96%
Yellow jacket wasp	85.61%

3.3 Discussion

The suggested pollinator classification algorithm results were dependent on several criteria. These have been uncovered and will have an impact on the classification outcome. The input image must meet certain requirements for it to produce the best and most accurate classification. The input image must be a clear image of the pollinator. This is the most critical factor to consider since if a blurry and unclear image is used as an input image, the algorithm will have difficulty extracting pollinator features as reported by [18].

Some limitations also existed from this initial study since this current work only relying on the dataset that have been collected manually by the researchers. One of them is that the number of pollinators displayed in the input image must be one since the algorithm is set up to classify the pollinator's type based on a single pollinator. If there are a lot of pollinators in one input image, the class's accuracy will decrease. Aside from that, it is preferable to have an input image with a white background. Because the algorithm depends on colours as one of the features to do classification, this criterion will also affect the value of feature extraction.

Finally, consider the angle of the input image. The object in the input image, which is the pollinator, must be at the best angle, which is flat facing from above. This is the optimal viewpoint because it allows the descriptors used in the feature extraction process to trace the shape and colour of the pollinator. A side-facing image should not be used as an input image since it will cause the descriptors to extract the value of the feature incorrectly.

As this work is aimed to be the basis for pollinator image classification, future work should be focus more on including more species and types of pollinators and employ a much more advanced algorithms that can classify any kind of pollinator images regardless of the limitations listed previously.

4. Conclusion

The proposed pollinator classification algorithm based on shape and colour features performs well and successfully classifies all the pollinators involved. Based on the features extraction method, area and colour features are beneficial in producing the best classification performance and accuracy. Aside from that, one of the factors that contribute to the system's success is the proper classifier used, which is Random Forest. Random Forest has given classification results with up to 90% accuracy for each type of pollinator.

One of the elements that contributed the most to the accuracy of the classification result is that the features used are appropriate. When the features selected are proper, it will make it easier to classify the pollinator because it gives a significant difference from one pollinator to another.

Overall, an algorithm that can automatically classify pollinator types could greatly assist people in expanding their understanding of pollinator types. It can be embedded in any related applications and help to save time and effort in searching for information on the pollinator. It is hoped that this study provides valuable insights and directions for future research in this field.

Acknowledgement
This research was not funded by any grant.

References
[1] Bugin, Giuseppe, Lucia Lenzi, Giulia Ranzani, Luigino Barisan, Claudio Porrini, Augusto Zanella, and Cristian Bolzonella. "Agriculture and Pollinating Insects, No Longer a Choice but a Need: EU Agriculture's Dependence on Pollinators in the 2007–2019 Period." *Sustainability* 14, no. 6 (2022): 3644. https://doi.org/10.3390/su14063644

[2] U.S. Department Of Agriculture. "Why is pollination important?" https://www.fs.usda.gov/managing-land/wildflowers/pollinators/importance

[3] Bell, Stephen AJ, Nigel Hunter, and Andrew Steed. "Lack of fire rather than pollinator absence may drive population decline in the critically endangered Banksia conferta (Proteaceae)." *Australian Journal of Botany* 70, no. 5 (2022): 372-383. https://doi.org/10.1071/BT21143

[4] Peixoto, Pedro Gomes, Heytor Lemos Martins, Bruna Cristina Pinto, Ana Luiza Franco, Larissa Souza Amaral, and Cristina Veloso de Castro. "The Significance of Pollination for Global Food Production and the Guarantee of Nutritional Security: A Literature Review." *Environmental Sciences Proceedings* 15, no. 1 (2022): 7. https://doi.org/10.3390/environsciproc2022015007

[5] Potts, Simon G., Jacobus C. Biesmeijer, Claire Kremen, Peter Neumann, Oliver Schweiger, and William E. Kunin. "Global pollinator declines: trends, impacts and drivers." *Trends in ecology & evolution* 25, no. 6 (2010): 345-353. https://doi.org/10.1016/j.tree.2010.01.007

[6] Bjerge, Kim, Hjalte MR Mann, and Toke Thomas Høye. "Real-time insect tracking and monitoring with computer vision and deep learning." *Remote Sensing in Ecology and Conservation* 8, no. 3 (2022): 315-327. https://doi.org/10.1002/rse2.245

[7] Ferreira, Alef Iury Siqueira, Nádia Felix Felipe da Silva, Fernanda Neiva Mesquita, Thierson Couto Rosa, Victor Hugo Monzón, and José Neiva Mesquita-Neto. "Automatic acoustic recognition of pollinating bee species can be highly

improved by Deep Learning models accompanied by pre-training and strong data augmentation." *Frontiers in Plant Science* 14 (2023): 1081050. https://doi.org/10.3389/fpls.2023.1081050

[8] Wäldchen, Jana, and Patrick Mäder. "Machine learning for image based species identification." *Methods in Ecology and Evolution* 9, no. 11 (2018): 2216-2225. https://doi.org/10.1111/2041-210X.13075

[9] Preethi, C., N. C. Brintha, and C. K. Yogesh. "An comprehensive survey on applications of precision agriculture in the context of weed classification, leave disease detection, yield prediction and UAV image analysis." *Advances in Parallel Computing Technologies and Applications* 40 (2021): 296. https://doi.org/10.3233/APC210152

[10] Nesteruk, Sergey, Dmitrii Shadrin, Mariia Pukalchik, Andrey Somov, Conrad Zeidler, Paul Zabel, and Daniel Schubert. "Image compression and plants classification using machine learning in controlled-environment agriculture: Antarctic station use case." *IEEE Sensors Journal* 21, no. 16 (2021): 17564-17572. https://doi.org/10.1109/JSEN.2021.3050084

[11] Deepan, P., and L. R. Sudha. "Object classification of remote sensing image using deep convolutional neural network." In *The cognitive approach in cloud computing and internet of things technologies for surveillance tracking systems*, pp. 107-120. Academic Press, 2020. https://doi.org/10.1016/B978-0-12-816385-6.00008-8

[12] Mustafa, Wan Azani, Nur Ain Alias, Mohd Aminuddin Jamlos, Shahrina Ismail, and Hiam Alquran. "A Recent Systematic Review of Cervical Cancer Diagnosis: Detection and Classification." *Journal of Advanced Research in Applied Sciences and Engineering Technology* (2022). https://doi.org/10.37934/araset.28.1.8196

[13] Tuda, Midori, and Alejandro Isabel Luna-Maldonado. "Image-based insect species and gender classification by trained supervised machine learning algorithms." *Ecological Informatics* 60 (2020): 101135. https://doi.org/10.1016/j.ecoinf.2020.101135

[14] Phasinam, Khongdet, and Thanwamas Kassanuk. "Machine learning and internet of things (IoT) for real-time image classification in smart agriculture." *ECS Transactions* 107, no. 1 (2022): 3305. https://doi.org/10.1149/10701.3305ecst

[15] Mustaffa, Mas Rina, Nyon Xin Yi, Lili Nurliyana Abdullah, and Nurul Amelina Nasharuddin. "Durian recognition based on multiple features and linear discriminant analysis." *Malaysian Journal of Computer Science* (2018): 57-72. https://doi.org/10.22452/mjcs.sp2018no1.5

[16] Russell, Stuart J., and Peter Norvig. "Artificial Intelligence A Modern Approach." (2010).

[17] Aldahiri, Amani, Bashair Alrashed, and Walayat Hussain. "Trends in using IoT with machine learning in health prediction system." *Forecasting* 3, no. 1 (2021): 181-206. https://doi.org/10.3390/forecast3010012

[18] Pathak, Harsh, C. Igathinathane, Kirk Howatt, and Zhao Zhang. "Machine learning and handcrafted image processing methods for classifying common weeds in corn field." *Smart Agricultural Technology* 5 (2023): 100249. https://doi.org/10.1016/j.atech.2023.100249

[19] Ebrahimi, M. A., Mohammad Hadi Khoshtaghaza, Saeid Minaei, and Bahareh Jamshidi. "Vision-based pest detection based on SVM classification method." *Computers and Electronics in Agriculture* 137 (2017): 52-58. https://doi.org/10.1016/j.compag.2017.03.016

[20] Zhong, Yuanhong, Junyuan Gao, Qilun Lei, and Yao Zhou. "A vision-based counting and recognition system for flying insects in intelligent agriculture." *Sensors* 18, no. 5 (2018): 1489. https://doi.org/10.3390/s18051489

[21] Fuchida, Masataka, Thejus Pathmakumar, Rajesh Elara Mohan, Ning Tan, and Akio Nakamura. "Vision-based perception and classification of mosquitoes using support vector machine." *Applied Sciences* 7, no. 1 (2017): 51. https://doi.org/10.3390/app7010051