

Short Communication: Insect detection using a machine learning model

SOMJIT HOMCHAN, YASH MUNNALAL GUPTA*

Department of Biology, Faculty of Science, Naresuan University, 99 Moo 9 Phitsanulok-Nakhonsawan Road, Phitsanulok 65000, Thailand.
Tel.: +66-55-963112, Fax.: +66-55-963113, *email: yashmunnalalg@nu.ac.th, yash.610@live.com

Manuscript received: 14 January 2021. Revision accepted: 3 February 2021.

Abstract. Homchan S, Gupta YM. 2021. Short Communication: Insect detection using a machine learning model. *Nusantara Bioscience* 13: 68-72. The key step in characterizing any organisms and their gender highly relies on correct identification of specimens. Here we aim to classify insect and their sex by supervised machine learning (ML) model. In the present preliminary study, we used a newly developed graphical user interface (GUI) based platform to create a machine learning model for classifying two economically important cricket species. This study aims to develop ML model for *Acheta domesticus* and *Gryllus bimaculatus* species classification and sexing. An experimental investigation was conducted to use Google teachable machine GTM for preliminary cricket species detection and sexing using pre-processed 2646 still images. An alternative method for image processing is used to extract still images from high-resolution video for optimum accuracy. Out of the 2646 images, 2247 were used for training ML model and 399 were used for testing the trained model. The prediction accuracy of trained model had 100 % accuracy to identify both species and their sex. The developed trained model can be integrated into the mobile application for cricket species classification and sexing. The present study may guide professionals in the field of life science to develop ML models based on image classification, and serve as an example for researchers and taxonomists to employ machine learning for species classification and sexing in the preliminary analysis. Apart from our main goals, the paper also intends to provide the possibility of ML models in biological studies and to conduct the preliminary assessment of biodiversity.

Keywords: Biodiversity, crickets, image recognition, insect, machine learning

INTRODUCTION

Classifying specimens/organisms is an initial and essential step for any biological studies. The scientific classification of a species demands prior knowledge for identification. However, automatic identification using machine learning model can be used to reduce time, cost, and labor. ML-based classifier is a rewarding tool identify organism and have been employed for insect identification (Yang et al. 2015). Various algorithms have been developed for insect classification (Ding and Taylor 2016; Kang et al. 2014), but all previously proposed methods are extremely dependent on algorithm development for optimal artificial neuron model for insect detection (Kaya and Kayci 2014). Nevertheless, these methods are also manual and time-consuming to develop insect detection algorithm (Larios et al. 2011; Kaya and Kayci 2014).

Artificial intelligence (AI) is revolutionizing human lifestyle and it has been integrated into many aspects (Deshpande and Kumar 2018). Machine learning (ML) is rapidly evolving and emerging in daily life. The bridge between users and developers is also expanding. It is essential to enlarge the involvement of professionals from different fields to apply their skills and develop a custom ML model without the need for coding/scripting skills. ML tools typically require technical skills that often include writing scripts to train classification models based on the input dataset. Particularly, professionals from life sciences

are usually unfamiliar with certain skills that require and align with machine learning tools, which typically access with scripting computer language. Although perverse studies have successfully demonstrated a specific insect classification system, it has certain limitations in terms of complex algorithm development process and usability in the real world. To date, various algorithms have been used and introduced for insect detection, but require scripting and algorithm testing to perform machine learning tasks (Kasinathan et al. 2020). Therefore, the present research concentrates on the investigation and usability of the graphical user interface systems to produce a machine learning model for biological images.

Google teachable machine (GTM) is a fairly new graphical user interface (GUI) based tool introduced in 2019 and it does not require coding or scripting skills to build ML models (Carney et al. 2020). GTM is based on TensorFlow, which is Javascript library of ML algorithms. Using the standalone TensorFlow library requires scripting skills in computer language like Python or JavaScript, which is eliminated in GTM and easy to access with GUI. Recently, GTM is being used for teaching purposes (Toivonen et al. 2020) and even has been used to design an educational program (Yu et al. 2019). ML algorithms have been applied for insect and crop pest classification (Tuda and Luna-Maldonado 2020; Ayan et al. 2020), but GTM has never been evaluated to develop ML model for insect classification purposes.

Insects are being consumed in several Asian countries including Thailand. These edible insects, especially cricket species are very popular in Thailand for a few decades (Hanboonsong et al. 2013). Edible insects are very popular among locals and easily recognizable due to their morphological features. However, it is difficult to distinguish closely related species. Previous studies of pest detection research had to deal with the image processing system. However, the classification model needed several technical steps (Miranda et al. 2014). The image processing for machine learning has evolved and can be done without scripting skills. Researchers have utilized GTM for classifying images (Ji and Jun 2020). Herein, we intended to employ GTM for preliminary cricket detection. Nevertheless, the same strategy can be adapted for pest detection and also can be integrated as a standalone application.

Machine learning has been adapted for numerous studies in life science. ML has been applied for insect detection using audio analysis (Silva et al. 2013), and also for the assessment of crop damage caused by insects using ML-driven drones (Puig et al. 2015). Even in November 2020, authors attempted ML approach to determine three insect species and their sex using still images (Tuda and Luna-Maldonado 2020). However, researchers also stated that an automated and online system yet to be developed. Moreover, determining and training model for ML proposes could be difficult, and integrating the ML-driven model requires skills that researchers from different filed often do not possess. Herein, we aim to use GTM for insect species and gender determination that does not require specific scripting or arithmetical skills. The GTM is a web-based machine learning system and also provides a script that can be integrated into mobile or web-based applications.

Crickets are economically very important species and widely consumed by local people in Thailand

(Gupta et al. 2020). Therefore, in the present preliminary study, we have used two edible cricket species to train machine learning model using GTM. Moreover, the sex-ratio is important in cricket breeding facilities due to the polyandrous nature of female crickets, the present ML model might serve as a preliminary tool to maintain sex-ratio in the captive population. The trained algorithm from the present study can be utilized for application development or can be accessed using web-link. To our knowledge, this is the very first attempt to use GTM for insect species detection and sex, which will also encourage researchers to use machine learning platforms like GTM.

MATERIALS AND METHODS

Data collection and pre-processing

Two cricket species, *Acheta domesticus* and *Gryllus bimaculatus* were collected from cricket breeding facilities. Initially, these crickets were scientifically identified and used by our research team for population genetic studies (Gupta et al. 2020). In the current application, females and males from both cricket species were used to capture

images from all directions. For the Image pre-processing, specimens were placed on white paper and 4K video at 30 frames per second (fps) was taken from all directions. Recorded 4K videos were used to extract frames for machine learning using Google Teachable Machine (GTM). GTM uses TensorFlow.js (Javascript library for machine learning). Extracted JPEG images were rectangular with the resolution of 500 dots per inch (DPI) having the dimension of 1280 X 720, JPEG images were separated into four different folders before uploading them into four classes on GTM server. Total dataset of 2,646 JPEG image files was created before using them for machine learning.

Training machine learning model on GTM

Pre-processed images were used and uploaded to GTM server. The training was repeated 100 times (epoch value). For deep learning with high accuracy, the batch size was set to 16 to divide data into smaller batches during training for each repetition. 85 % of the dataset was used to train the model and the other 15% of dataset was never been used for training but used to check the performance and accuracy of the trained model. Previously, the randomized splitting method was also used to divide the pest image dataset prior to machine learning (Ayan et al. 2020).

To evaluate the model, the advanced parameters were used at GTM platform to compute accuracy per cricket species class, accuracy per repetition, loss per repetition, and to generate confusion matrix.

Sample data and JavaScript availability:

<https://github.com/yashmgupta/Insect-species-detection->

RESULTS AND DISCUSSION

Firstly, to increase the accuracy and capture the image from all directions, 4K (around 4000 pixels in the horizontal direction) video was taken and the process to obtain JPEG files of still images having a resolution of 500 DPI with 24-bit depth. Recently, researchers used a scanning method to obtain 600 JPEG files of insects (Tuda and Luna-Maldonado 2020). In the present study, we used an alternative and less time-consuming method to generate pre-processed image files for machine learning purposes. Moreover, scanning insect specimens cannot capture the visual of a specimen from all directions that might affect the accuracy of the trained model for species or sex determination using a machine learning model. Usually, the train model accuracy increases by increasing the numbers and the quality of images (Ji and Jun 2020). To reduce time consumption for data pre-processing and increase the probability of training machine learning model with the highest accuracy, we developed an alternative method, which resulted in 2646 pre-pressed image files for deep machine learning. Hereto, 2247 images were used for training the model with GTM and 399 images were used as test samples to check the performance of the trained model.

The results show that the trained model has a 100% probability for categorizing cricket images to correct species and sex. The 399 cricket images were employed to check accuracy. There was no instance where a single cricket image was classified for more than one insect species or sex. In the present supervised machine learning test, for all four classes, the accuracy was one (1.00) as shown in Table 1. This means all samples used to test the trained model were successfully classified according to their features (species and sex) Figure 1.

Default options of GTM for epochs are fifty, but in present supervised machine learning model, the whole dataset went through the training a hundred times for better prediction and accuracy (as mentioned in method section). The training and testing accuracy per epoch is shown in Figure 2 (A). Both accuracies reached their threshold after only seven epochs of training with the whole dataset. In the present machine learning model, the prediction for correct cricket species and sex becomes practically perfect after the 7th epoch, as loss per each became zero at the 20th epoch and remained stagnant till the last epoch, Figure 2 (B). When the prediction is perfect, the accuracy becomes one, and after that prediction, if the confidence is hundred percent, the loss becomes zero (Carney et al. 2020). The validation result matches and confirms that our ML model learned appropriately without underfit or overfit.

As described in the method section, we extracted still images from high-resolution video to be converted into still images for training ML model to capture morphological keys of cricket species. Recently, researchers have used scanners for image processing (Tuda and Luna-Maldonado 2020), but we have altered the conventional method of pre-processing still images for classification. The ML model based on image classifications has been applied to map tree species (Wang et al. 2018; Lim et al. 2019), and using object-based analysis to classify deciduous tree species (Franklin and Ahmed 2018).

Insects, including crickets, are relatively small, which often makes them harder to identify or determine their sex without experience or expertise. However, the most common visible feature in sexing for cricket is the presence of an ovipositor in female crickets. Our application of GTM for primary identification based on image classification confirms that the machine learning model trained for cricket species identification and sexing. Previously, researchers suggested applying molecular biology techniques for sustainable insect breeding for food and feed (Gupta and Homchan 2019). Herein, we are suggesting that the developed ML model can also be applied to maintain sex-ratio in the breeding facility.

Besides, this study is the first step towards educating researchers to use machine learning from GTM platform in biological sciences. Result of the trained model for cricket species classification using GTM is encouraging. The developed algorithm can be excess via an online platform, or can be used with a mobile application (developed to run ML model in the real environment). This research was concerned with the usability of GTM for research purposes;

however, the findings suggest that this approach can be useful for developing algorithm using an online platform for insect detection with direct practical relevance. Moreover, TensorFlow is huge library of machine learning algorithms extensively used for research purposing, including drug discovery and image classification (Abadi et al. 2016). GTM uses pretrained neural network of mobilenet models from TensorFlow library, which has separate 28 layers for the image classification with high accuracy (Howard et al. 2017). The finding also has important implications for assisting active sample collectors without prior knowledge about cricket or any other insect species for assessment of biodiversity or population analysis.

The proposed method can be adapted for several biological science applications and can be readily used in practice for preliminary work. In the present study, particular attention is paid to develop an adaptable approach for ML practice in biological science where observation-based assessment is needed (e.g. on-site cricket sample collection). On the basis of the promising results of two cricket species classification and sexing, the ML model with other cricket species will be developed and investigated in future research.

Table 1. Accuracy per class of cricket species and their gender.

Classes	Accuracy	Test samples
ADF	1.00	118
ADM	1.00	118
GBF	1.00	78
GBM	1.00	85

Note: ADF: *Acheta domesticus* female, ADM: *Acheta domesticus* male, GBF: *Gryllus bimaculatus* female, GBM: *Gryllus bimaculatus* male

Classes	ADF	118	0	0	0
	ADM	0	118	0	0
	GBF	0	0	78	0
	GBM	0	0	0	85
		ADF	ADM	GBF	GBM
		Prediction			

Figure 1. Confusion matrix of classes according to the provided samples (Y-axis) and predicted classes based on the trained model (X-axis). The trained model showed no misclassification for 399 samples that new-before-seen by ML model.

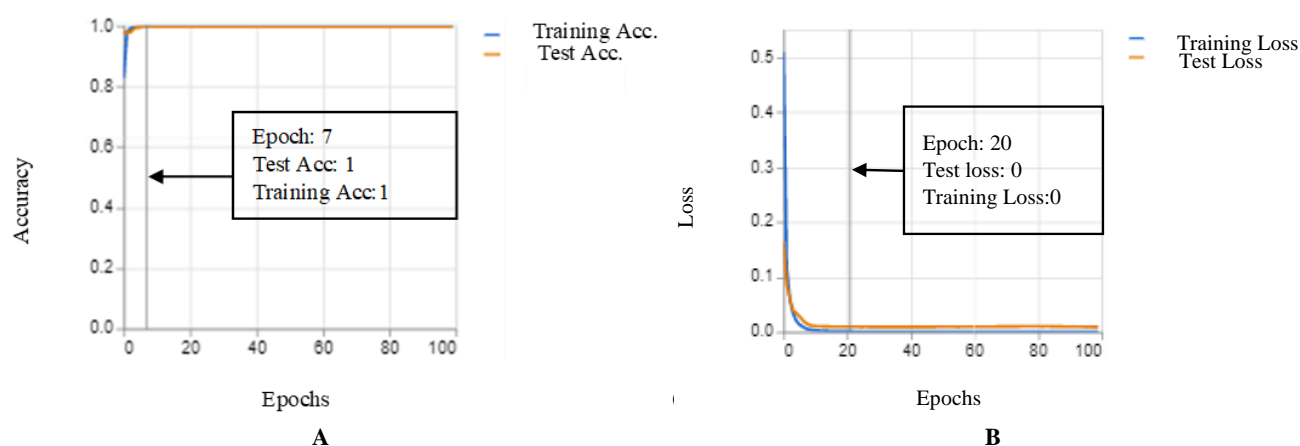


Figure 2. Plots for training and testing dataset result validation: A. Plot of training and test accuracy against epochs. B. The plot of training loss and test loss against epochs

The present study was designed to determine the usability of machine learning platform like GTM for preliminary species detection and sexing. The result of the investigation shows that the ML model can be developed with ease using GTM and incorporated in mobile and website or can be directly access with the shareable link. The key strengths of this study are its user-friendly online ML platform and accessibility after training the cricket classification model. Although the current ML model is based on two cricket species, the findings suggest that ML model can be developed for classification of several cricket species with different morphological features. A reasonable approach to capture details is extracting still images from the video; therefore, we suggest a similar to this one should carry out before building the ML model. Taken together, our results suggest that ML model can be trained on the GTM platform for image-based classification in biological sciences, including insect classification.

ACKNOWLEDGEMENTS

We would like to thank Ms. Satwika Nindya Kirana for polishing English and giving valuable comments. This research has been supported by Department of Biology, Faculty of Science, Naresuan University, Thailand.

REFERENCES

- Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mane D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viegas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng Y. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. Arxiv:1603.04467.
- Ayan E, Erbay H, Varçın F. 2020. Crop pest classification with a genetic algorithm-based weighted ensemble of deep convolutional neural networks. *Comput Electron Agric* 179: 105809. DOI: 10.1016/j.compag.2020.105809.
- Carney M, Webster B, Alvarado I, Phillips K, Howell N, Griffith J, Jongejan J, Pitaru A, Chen A. 2020. Teachable Machine: Approachable web-based tool for exploring machine learning classification. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu, 25-30 April 2020. [USA] DOI: 10.1145/3334480.3382839.
- Ding W, Taylor G. 2016. Automatic moth detection from trap images for pest management. *Comput Electron Agric* 123: 17-28. DOI: 10.1016/j.compag.2016.02.003.
- Deshpande A, Kumar M 2018. *Artificial Intelligence for Big Data: Complete Guide to Automating Big Data Solutions Using Artificial Intelligence Techniques*. Packt Publishing Ltd., Birmingham, UK.
- Franklin SE, Ahmed OS. 2018. Deciduous tree species classification using object-based analysis and machine learning with unmanned aerial vehicle multispectral data. *Int J Remote Sens* 39(15-16): 5236-5245. DOI: 10.1080/01431161.2017.1363442.
- Gupta YM, Homchan S. 2019. Preventive measures and sustainable practices for insect breeding. *J Biol Sci* 5 (4): 1-4.
- Gupta YM, Tanasarnpaiboon S, Buddhachat K, Peyachoknagul S, Inthim P, Homchan S. 2020. Development of microsatellite markers for the house cricket, *Acheta domestica* (Orthoptera: Gryllidae). *Biodiversitas* 21(9): 4094-4099. DOI: 10.13057/biodiv/d210921.
- Hanboonsong Y, Jamjanya T, Durst PB. 2013. *Six-Legged Livestock: Edible Insect Farming, Collection and Marketing In Thailand*. FAO Regional Office for Asia and the Pacific, Bangkok.
- Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. Arxiv:1704.04861.
- Ji SY, Jun HJ. 2020. Deep learning model for form recognition and structural member classification of east Asian traditional buildings. *Sustainability* 12(13): 5292. DOI: 10.3390/su12135292.
- Kang SH, Cho JH, Lee SH. 2014. Identification of butterflies based on their shapes when viewed from different angles using an artificial neural network. *J Asia Pac Entomol* 17(2): 143-149. DOI: 10.1016/j.aspen.2013.12.004.
- Kasinathan T, Singaraju D, Uyyala SR. 2020. Insect classification and detection in field crops using modern machine learning techniques. *Inform Process Agric*. DOI: 10.1016/j.inpa.2020.09.006.
- Kaya Y, Kayci L. 2014. Application of artificial neural network for automatic detection of butterfly species using color and texture features. *Vis Comput* 30: 71-79. DOI: 10.1007/s00371-013-0782-8.
- Larios N, Lin J, Zhang M, Lytle D, Moldenke A, Shapiro L, Dietterich T. 2011. Stacked spatial-pyramid kernel: An object-class recognition method to combine scores from random trees. *IEEE Workshop on Applications of Computer Vision (WACV)* 329-335. DOI: 10.1.1.363.9133.
- Lim J, Kim KM, Jin R. 2019. Tree species classification using Hyperion and sentinel-2 data with machine learning in South Korea and China. *ISPRS Int J Geoinf* 8(3): 150-177. DOI: 10.3390/ijgi8030150.

- Miranda JL, Gerardo BD, Tanguilig Iii BT. 2014. Pest detection and extraction using image processing techniques. *Int J Comput Commun Eng* 3(3): 189-192. DOI: 10.7763/IJCC.2014.V3.317.
- Puig E, Gonzalez F, Hamilton G, Grundy P. 2015. Assessment of crop insect damage using unmanned aerial systems: A machine learning approach. *Proc. 21st Int. Congr. Modelling Simulation*, Gold Coast, Australia, 29 November - 4 December 2015. DOI: 10.36334/modsim.2015.f12.puig.
- Silva DF, De Souza VM, Batista GE, Keogh E, Ellis DP. 2013. Applying machine learning and audio analysis techniques to insect recognition in intelligent traps. *12th International Conference on Machine Learning and Applications*, 99-104 2013. DOI: 10.1109/icmla.2013.24.
- Toivonen T, Jormanainen I, Kahila J, Tedre M, Valtonen T, Vartiainen H. 2020. Co-designing machine learning apps in k-12 with primary school children. *IEEE 20th International Conference on Advanced Learning Technologies (ICALT)*, 308-310. DOI: 10.1109/ICALT49669.2020.00099.
- Tuda M, Luna-Maldonado AI. 2020. Image-based insect species and gender classification by trained supervised machine learning algorithms. *Ecol Inform* 60: 101135. DOI: 10.1016/j.ecoinf.2020.101135.
- Wang D, Wan B, Qiu P, Su Y, Guo Q, Wu X. 2018. Artificial mangrove species mapping using pléiades-1: An evaluation of pixel-based and object-based classifications with selected machine learning algorithms. *Remote Sens* 10: 294-315. DOI: 10.3390/rs10020294.
- Yang HP, Ma CS, Wen H, Zhan QB, Wang XL. 2015. A tool for developing an automatic insect identification system based on wing outlines. *Sci Rep* 5: 1-11. DOI: 10.1038/srep12786.
- Yu WJ, Jang JH, Ahn JM, Park DR, Yoo IH, Bae YK, Kim WY. 2019. Design of artificial intelligence education program based on design-based research. *Int J Adv Smart Converg* 8: 113-120. DOI: 10.7236/IJASC.2019.8.4.113.