Prediction of a plant intracellular metabolite content class using image-based deep learning

Neeraja M Krishnan and Binay Panda*

Ganit Labs Foundation, India

***Correspondence: Binay Panda (binay@ganitlabs.in)**

Abstract

Plant-derived secondary metabolites play a vital role in the food, pharmaceutical, agrochemical, and cosmetic industry. Metabolite concentrations are measured after an elaborate process of extraction, biochemistry, and analyses, that requires time, access to expensive equipment, reagents and specialized skills. Additionally, the concentration of secondary metabolites often varies widely among plants, even within a small location. A quick method to estimate the metabolite concentration class (high or low) will significantly help in selecting trees yielding high metabolites and use those trees to collect parts to be used in the metabolite production process. There are no methods to do this currently. Here, we demonstrate a deep learning approach to estimate the intracellular concentration class of a metabolite, azadirachtin, using models built with images of leaves and fruits collected from randomly selected Azadirachta indica (neem) trees in an area spanning >500,000 sq km and their corresponding biochemically measured metabolite concentrations. We divided the input data randomly into training- and test-sets ten times to avoid sampling bias and to optimize the model parameters during cross-validation. The training-set (background-removed images augmented with seven synthetic distortions) contained >83,000 fruit and >86,000 leaf images. The best models yielded prediction errors of 19.13% and 15.11% (for fruit), and 8% and 26.67% (for leaf), each, for low and high metabolite classes, respectively. We further validated the fruit model using independently collected fruit images from different locations spanning nearly 130,000 sq km, with 70% accuracy. We developed a desktop application to scan off-line image(s) and a mobile application for real-time utility (while fruits and leaves were still attached to the trees) to predict the metabolite class. Our work is the first demonstration of deep learning to estimate an intracellular metabolite class using images and has broad applications and utility.

Introduction

Measuring the concentration of an analyte such as metabolite, enzyme, protein or any other chemical moiety within plants, animals and microbial cells is a frequent practice in biology. To do this, chemical, biochemical, immunological or imaging-based methods are followed, where each method provides a different type of readout. Although accurate and precise, the currently employed analytical methods require extensive sample handling and preparation time, expensive reagents and equipment, and specialized skills. In circumstances where knowing the exact amount of the intracellular metabolite is not necessary, and a quick and rough estimate of its class (either high or low) is enough, use of conventional measurement methods although

unnecessary is currently unavoidable. A method that provides a quick readout of the analyte class, however ideal in such cases, is currently not available.

Metabolites, primary and secondary, are intermediate products of metabolic reactions catalyzed by enzymes. Examples of some primary metabolites are amino acids, vitamins, organic acids, and nucleotides. Living cells synthesize most of the primary metabolites as they are essential for their growth. Secondary metabolites, especially those derived from microbial and plant sources, such as drugs, fragrances, dyes, pigments, pesticides, and food additives, are not required for the cell's primary metabolic processes but have a wide range of application in agriculture and pharmaceutical industry. The commercial importance of secondary metabolites has invoked the exploration of producing bioactive plant metabolites synthetically in the laboratory, either by total synthesis, plant tissue culture or metabolic engineering (Keasling 2014).

Machine learning and artificial intelligence-based methods have recently found their utility in biology (Minervini et al. 2014; Rostam et al. 2017). Algorithms based on deep learning can predict patterns from raw biological data (genomic data, images, proteins, and other analytes) and help to generate biological inferences. Deep learning methods outperform conventional methods of image classification, such as principal components identification, and their usage in Support Vector Machines and Random Forest classifiers (Kovalev et al. 2017). They are, therefore, more suitable for image-based analyses.

In the current manuscript, we focussed on a commercially important secondary metabolite, azadirachtin, a potent and versatile tetranotriterpenoid found predominantly in the fruits and to a lesser degree, in the leaves of the tree, Azadirachta indica (neem) (Schmutterer 2002). We used an image-based deep learning approach utilizing convolutional neural networks (CNNs) to learn from the images of leaves and fruits from sampled neem trees, alongside their measured concentrations of azadirachtin to build models that can predict the metabolite concentration class for an incoming new leaf or fruit.

Materials and Methods

The flow diagram in **Figure 1** describes the training set and training criteria used in building models, and the different thresholds for azadirachtin class (azaL – low azadirachtin and azaH – high azadirachtin) prediction accuracies used in short-listing and finalizing models.

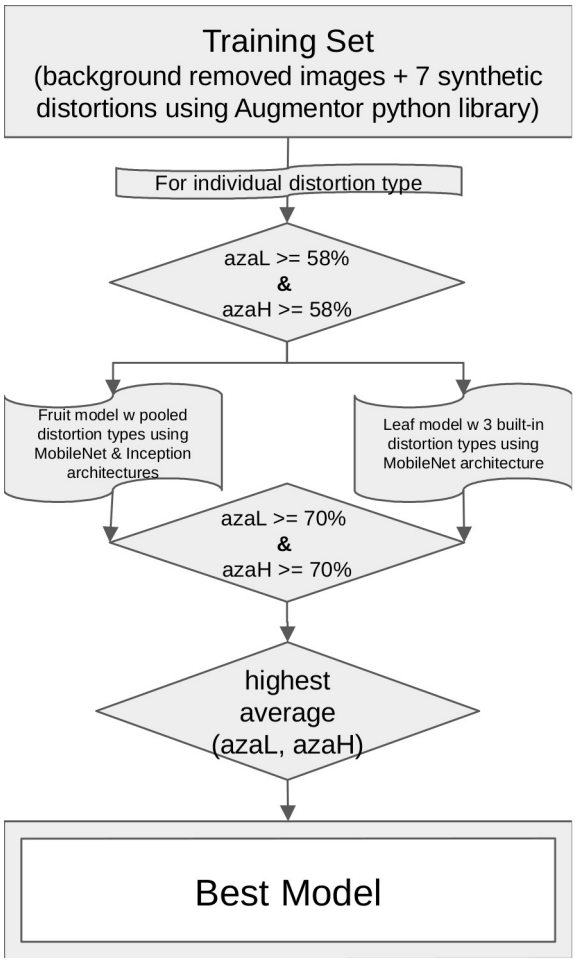

Figure 1: *Flow diagram describing the training set and training criteria used in building models, and the different thresholds for azadirachtin class (azaL – low azadirachtin and azaH – high azadirachtin) prediction accuracies used in short-listing and finalizing models.*

Plant material collection and Metabolite measurement

We randomly sampled neem trees in an area covering >500,000 sq km and collected five leaves and fruits per tree. Both leaves and fruits were photographed in controlled light conditions. This minimized variability during image capture. Azadirachtin (azadirachtin A and azadirachtin B, together) was measured using reverse-phase columns on High Performance Liquid Chromatography (HPLC) instrument with azadirachtin (A+B) analytical standards (EID Parry, India).

Augmentation of images

The original images were first subjected to background removal using an online editing tool named PhotoScissors (https://online.photoscissors.com/). This step was performed to base the classification on the foreground, and not the variable background (white in case of the training set). A Python module, Augmentor (Bloice et al. 2017), was used to augment the background-removed images with 7 types of distortions: Rotate [rotate (probability=1, max_left_rotation =1..25, max_right_rotation=1..25), rotate90 and rotate270],

Zoom [percentage_area=0.8 or min_factor =1.1 and max_factor=1.5], Flip [left-right or top-bottom], Distortion [random (grid_width=4, grid_height =4, magnitude=8) or gaussian (grid_width=5, grid_height=7, magnitude=2, corner="bell", method="in" or "out", mex=0.5, mey=0.5, sdx=0.05, sdy=0.05)], Crop [random (percentage_area=0.5) or by_size (width=5, 10, 15, 20, 25, 30, 35 or 40, height=6.25, 12.5, 18.75, 25, 31.25, 37.5, 43.75 or 50, centre=True) or crop_centre (percentage_area=0.8, randomise_percentage_area=False)], Brightness [random (min_factor=0.5, max_factor=1.6)] and Resize [width = 60..220, height = 100..375]. Representative original images of leaves and fruits for azaL and azaH classes and post distortions are shown in **Figure 2**.

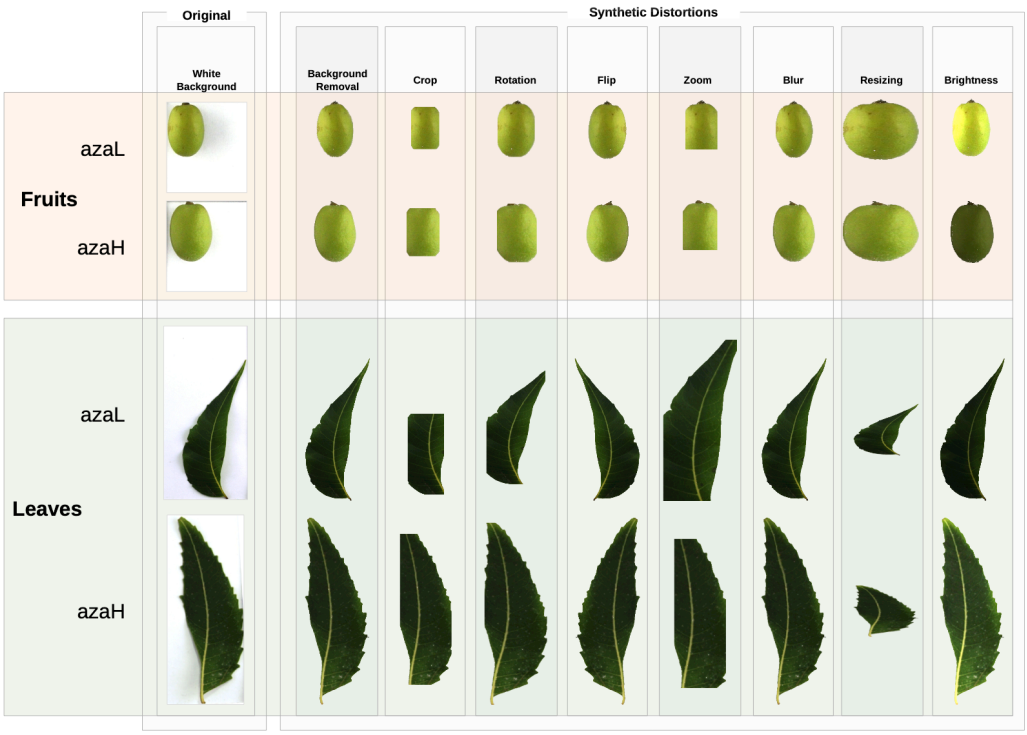

Figure 2: *Representative images of neem leaves and fruits for azaL and azaH classes before and after synthetic distortions.*

Learning and Test Set Classification

As azadirachtin derived from seed kernel is considered to be the gold standard test, we used seed kernel-derived metabolite values to set metabolite class as either “low (azaL)” or “high (azaH)”. The minimum and maximum were determined for the azadirachtin values. The average (avg) was determined as the mid-point of the min-max extremes. An azadirachtin class was set as ‘low’ and ‘high’ when the measured concentrations of the metabolite were below avg - 0.1avg and above avg + 0.1avg, respectively. Accordingly, an azadirachtin value above 0.65% was determined as ‘high’, and a value below 0.53% was determined as ‘low’. The sample numbers for the azaL and azaH classes in ten learning and test sets randomly generated for 10-fold cross validation are summarised in **Table 1**. The corresponding distorted images were also divided into learning and test sets for training and classification.

Algorithm and Framework

We named the tool 'Aza Classifier' based on the test case, herein, to classify azadirachtin content class. Aza Classifer (desktop version) is developed using JAVA swing GUI (Horstmann & Cornell 2013) for its native look and is backed up with Linux shell scripting and Python3 in its backend. The desktop version of Aza Classifier was built to classify azadirachtin content class (low or high) for a single leaf or fruit image, or multiple images saved in a folder (**Figure 3A**). The results from the desktop app were bulk exported into an Excel sheet while storing the image identity, with predicted azadirachtin class, and prediction confidence (probability ranging from 0 to 1) and analyzed further. We used TensorFlow 1.6.0 packages (Abadi et al. 2016), an open-source framework for building deep learning neural networks (Schmidhuber 2015), in Python3 to build predictive models that could classify images of leaves and fruits. There are a few challenges in visual recognition of images, particularly, for on-device or embedded applications to run the models effectively in an environment where resources like computation, power, and space are limited. The MobileNet architecture (Howard et al. 2017) provided a solution for this. It is a family of mobile-first computer vision models for TensorFlow, designed specifically for an effective increase in accuracy, with limited resources for an on-device or embedded application. MobileNets are actually small, low-power, low-latency models designed to deal with the limited resources of a number of use cases. The latest release (1.11) consists of a model definition for MobileNets in TensorFlow using TF-Slim in addition to 16 pre-trained ImageNet classification checkpoints for use in mobile projects of any size. The mobile version of Aza Classifer (Android app) was built using Android Studio 3.0 (Smyth 2017), compiled and run based on TensorFlow Android version 1.6.0. The android app allows a live or real-time classification of the azadirachtin class by pointing the camera of the mobile device to a leaf or a fruit attached to a tree (**Figure 3B**), to predict its metabolite content class. A "Save" option in the app enables the user to capture an image along with the azadirachtin classification while providing an identifier to the captured image. In cases of fluctuation in the image capture process, the user has the provision to capture the same image multiple times and choose the most common readout using the app.

Training set formation and model testing using TensorFlow library with CNN architecture

Using TensorFlow, we built azadirachtin prediction models for 17 types of configurations including 16 Mobilenet (pre-trained deep CNN network versions for width multipliers of 0.25, 0.50, 0.75 and 1.0, and image input sizes of 128, 160, 192, and 224) and Inception (pre-trained deep CNN network) v3 (Szegedy et al. 2015) models. Each model was trained for 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500 and 5000 training steps.

Images were divided into azaL and azaH classes. The training set was formed using the images after they were subjected to background removal, and then further to the 7 distortions externally (as described earlier). Additionally, for leaf data, 3 distortions to the background-removed images made internally using TensorFlow, was used for training. The test sets in both cases were the set of images not used during training.

For each model, the azaL and azaH classification sensitivities were average across the ten test sets. Presence of at least one model with >58% classification sensitivity for both azaL and azaH classes was ascertained for each individual distortion type. Following this ascertainment, 160 MobileNet and 10 Inception models (with the ten training steps mentioned previously) were constructed after training with all the 8 types of images together and tested with images not used during training. For the leaf, the background-removed images were also distorted internally (Crop - 30%, Scale - 30% and Flip - left and right) using the python module 'retrain' of TensorFlow library (https://github.com/tensorflow/hub/blob/master/examples/image_retraining/retrain.py), followed by training and classification using 160 MobileNet models. The fruit models from MobileNet and Inception architectures, with at least 80% classification sensitivities for both azaL and azaH classes were shortlisted. Likewise, the leaf models were shortlisted from TensorFlow's inbuilt distortions while retraining using MobileNet, with at least 70% classification sensitivities for both classes. These thresholds were set as the highest common classification sensitivities for azaL and azaH classes. The final fruit and leaf models were chosen with the highest average classification sensitivity between azaL and azaH classes.

Independent validation of the model

Images of randomly sampled fruits from independent locations covering nearly 130,000 sq kms, which were not used for training or testing, served as a validation set.

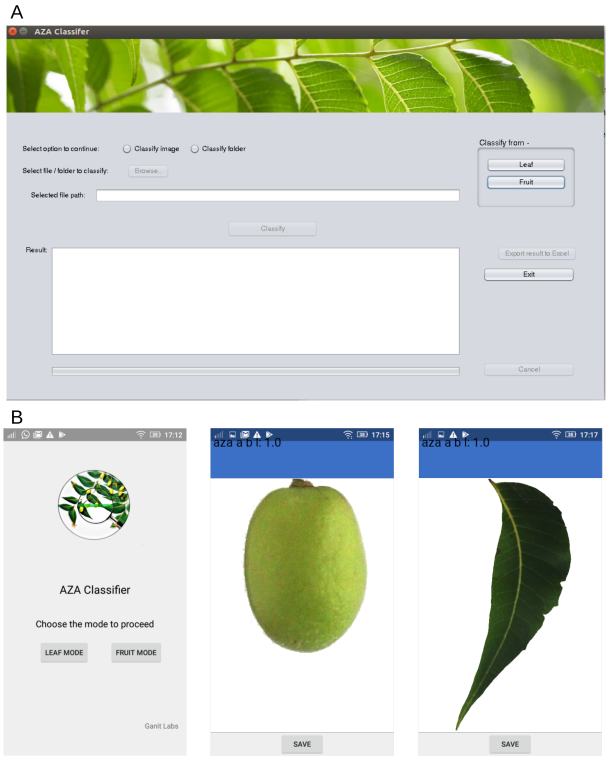

Figure 3: *Interface of the classifier app in the A. Desktop and B. Mobile versions.*

Results

Augmentation of background-removed images by synthetic distortions

We used seven synthetic distortions (as described in the Methods section) on the background-removed images. The number of images used in training and testing for each distortion type are listed in **Table 2**. The image numbers were highest for the Rotation distortion types augmented ~60-fold over the original *i.e.* background-removed images, followed by ~22-fold for the Crop distortion and ~17-fold for the Resize distortion type. Pooling all synthetic distortions, the numbers of images used for training were 83,898 (48285 for azaL & 35613 for azaH) for fruits and 86,311 (51,495 for azaL & 34,816 for azaH) for leaves (**Table 2**).

Suitable fruit and leaf models for individual distortion types using MobileNet architecture

Fruit and leaf models for individual distortion types that passed at least 58% classification sensitivity for azaL and azaH content classes were deemed suitable **(Supplementary Table S1)**. Among the fruit models, there were 34 suitable models for Brightness, 35 for Resize, 21 for Crop, 47 for Random/Gaussian blur, 49 for Rotate, 40 for background-removed, 49 for Zoom and 53 for Flip, and among the leaf models, there were 13 suitable ones for Brightness, 15 for Resize, 4 for Crop, 15 for Random/Gaussian blur, 1 for Rotate, 21 for background-removed, 24 for Zoom, and 16 for Flip distortion types. This criterion was used to decide on the pooling of images from all distortion types, for training and classification, to construct near-final models.

Suitable fruit and leaf models for the second training set comprising all background-removed images including all distortions post background-removal, using MobileNet and Inception architectures

Fruit and leaf models for all the distortion types were deemed suitable if they predicted with at least 70% classification sensitivity for azaL and azaH content classes **(Supplementary Tables S2-S3)**. As per these criteria, 15 fruit models were deemed suitable with the MobileNet architecture, and 7 with the Inception architecture. No leaf model passed these criteria, with either architectures.

Suitable leaf models for background-removed images, factoring in three in built distortions using TensorFlow library, with MobileNet architecture

Six leaf models were deemed suitable based on at least 70% classification sensitivity for azaL and azaH content classes within the original background-removed test set of leaf images **(Supplementary Table S4)**.

Final fruit and leaf models of azadirachtin classification

The thresholds for shortlisting fruit and leaf models trained with pooled augmented data were individually set to be the highest common classification sensitivities for azaL and azaH classes, at 80% and 70%, respectively. Two fruit models and six leaf models passed these criteria. The final models were chosen from Inception V3 architecture, at 2500 training steps, for the fruit, with an average classification sensitivity of 82.88%, and from the MobileNet

architecture, with built-in distortions while retraining, image resolution of 128, width multiplier of 0.50, at 2500 training steps, for the leaf, with an average classification sensitivity of 82.67% **(Table 3)**. Since the leaf model was derived from a training set augmented with only three distortion types (Brightness, Flip, Gaussian/Random blur and Zoom), it was less sensitive to classifying azadirachtin content within a test set that accommodated eight distortion types, including background-removal, thus resulting in a slightly lower classification sensitivity of >70% for both azaL and azaH classes **(Table 3)**. For Resize, Rotate and Crop distortion types, while the azaL classification sensitivity was 90.67%, 77.26% and 89.27%, respectively, the azaH classification sensitivity remained below 50% at 25.74%, 42.08% and 42.58%, respectively **(Table 3)**.

Independent validation of fruit models

The final fruit model was validated in an independent set of fruit images, with ~70% classification sensitivity, respectively **(Table 4)**.

Discussion

Deep learning methods are producing impressive results across several domains, especially, visual and auditory recognition (Lecun et al. 2015). Data-intensive biological problems are well suited to be solved using deep learning methods (Ching et al. 2018). Biologically inspired neural networks are a class of machine learning algorithms that enable learning from data. Deep learning requires a neural network with multiple layers. Based on the nature of data and the type of question being asked, deep learning methods use either supervised, or unsupervised, or reinforcement learning-based training models. CNNs or ConvNets are multi-layered neural networks trained with the back-propagation algorithm, used in recognizing images with minimum pre-processing. In 1998, LeCun and coworkers had described a machine learning technique that was built to learn automatically with less hand-designed heuristics (LeCun et al. 1998). This formed the basis for the development of the CNN field. CNNs combine three architectural ideas: local receptive fields shared weights and, sometimes, spatial or temporal sub-sampling (LeCun and Bengio 1995). Biological data is often complex, multi-dimensional and heterogeneous. However, the possibility of using deep learning methods to discover patterns in such large and complex biological datasets is promising. Till date, image-based neural networks have been used to find patterns that aid disease detection and diagnosis (Esteva et al. 2017, Gulshan et al. 2016, Lakhani and Sundaram 2017), secondary structure prediction in proteins (Wang et al. 2017, Spencer et al. 2015, Wang et al. 2016), effects of single nucleotide variation in genes (Quang et al. 2015), plant phenotyping (Pound et al. 2017) and other biological applications.

Plant-based metabolites offer a cheaper, non-toxic and environment-friendly alternative to chemical pesticides. Azadirachtin, derived from the fruits of the tree, *Azadirachta indica* (neem), is one such metabolite with known anti-feedant and insect-repellant properties (Schmutterer et al. 2017). Seed kernels derived from the tree's fruits are the primary source of commercial azadirachtin. Although they are an inexpensive source of the metabolite, the current process of extraction is not scalable due to limited seed availability in the market. The commercial source of seeds is wild trees, and the number of

trees in the wild is limited. Additionally, azadirachtin content varies a lot depending on the geography of the tree locations, quality of the seeds, age of the trees, soil conditions and climatic factors (Chary 2011; Kaushik et al. 2007; Sidhu et al. 2003; Singh et al. 1999; Srivastava et al. 2010), thus causing the commercially available seeds to yield varying levels of azadirachtin. A previous study has reported high variation, as high as 6-fold, in azadirachtin-A and azadirachtin-B content across seed kernels (Sidhu et al. 2003). Currently, there are no methods to select high metabolite containing fruits. To circumvent this, alternatives to plant-based azadirachtin were also investigated in the past. For example, Veitch et al. (Veitch et al. 2007) attempted total organic synthesis of azadirachtin. However, the process was time-consuming with a low yield (Veitch et al. 2008). Recent successful advances in synthetic biology such as producing isoprenoids like β-farnesene and the antimalarial drug, artemisinin, in heterologous hosts (Keasling 2014; Meadows et al. 2016; Paddon et al. 2013; Ro et al. 2006) have resurrected the hope for bulk production of metabolites, independent of their natural source. However, synthetic biology requires full knowledge of the metabolite synthesis pathway, and optimization of the pathway and production process within the heterologous host(s). Until a complete synthetic biology solution emerges, identifying high azadirachtin-bearing fruits and selecting them as commercial sources will greatly boost the production of high-grade azadirachtin.

Current methods for metabolite screening, although precise, are cumbersome and require a laboratory equipped with expensive equipment, reagents, and expert personnel. Additionally, these methods require all collected fruits to be transported to a central laboratory, thus making the process lengthy and costly. Often, a method that provides a less precise readout of the azadirachtin level, for e.g., low or high, is sufficient to make an initial choice before commercial extraction. The current study is a step towards this aim, and provides a simple, inexpensive and on-the-field method that can be used by anyone, even with little or no training at the collection sites to selectively choose high azadirachtin yielding fruits.

India has nearly 337 million smartphone users, which is expected to grow to 490 million by 2022 (eMarketer survey May 2018). The country is also one of the fastest growing markets for mobile device-based delivery platforms for services and e-commerce. Therefore, we explored the possibility of devising a mobile application that can be used at the collection site to determine whether a plant organ has high or low metabolite content. The potential of such an application based on deep learning technology to classify images and predict biological measurements is very enticing. In the current report, we show that it is possible to use an image of a fruit or a leaf to determine whether the intracellular concentration of an analyte within the corresponding cells is high or low. Our method provided us with nearly 70% validation with the fruit model when trained with image distortions. Since the seed kernels from fruits are the primary source of azadirachtin, our model is well suited to pre-select fruits that can be used as commercial sources. There are exciting possibilities that can be tested based on the similar concept demonstrated here. For example, one can use images for quality control in the fruit industry or deciding the food content class, for example, sweetness, bitterness, in fruits or vegetables. There are other requirments as well in the horticultural/agricultural industry

that can use image-based classification to deduce characteristics of a certain trait.

The prediction accuracy of our classifier app was not 100%, and we got errors of 19.13% and 15.11%, and 8% and 26.67%, each, using the fruit and leaf models, respectively, for the low and high metabolite class, respectively. Although our initial models are promising for image-based metabolite class prediction, there are several limitations to our study. The performance of any machine learning approach is dependent on both, adequacy of the sampled data and the complexity of the underlying model. In cases where the training set data is limiting, chances of the model over-fitting to the training data and not generalizing to other 'unseen' data are greater, while in cases where the model is over-simplified or over-regularized, chances of the model under-fitting the data are greater. The first limitation of our study is the relatively narrow size of the training set and lack of diversity, in terms of brightness, camera magnification, blur, left-right flip, top-bottom flip, degree of rotation, etc, Therefore, it is possible that the model might be an over-fit, and not be a suitable predictor for unseen data. Although we have attempted to lower the impact of the first limitation by augmenting the training set with synthetic distortions, the prediction accuracy of Aza Classifier is expected to increase with naturally augmenting the training set, in both size and diversity, via further sampling, imaging and metabolite content analyses of neem leaves and fruits. In doing so, the aim would first be to utilize the additional data to validate the underlying model parameters, and second, to enhance the training set, in an iterative manner to minimize both, cross-validation and true validation errors. The second limitation of our study is the lack of metabolite data from individual fruits and leaves of a tree. Although we collected 10 samples per tree (5 fruits and 5 leaves), we had single metabolite readout per plant part. Results from methods that can measure metabolite from a small amount of tissue (such as a single fruit or leaf), can be variable with such a low amount of tissue. With a reasonable assumption that the intra-tree variation of metabolite content across fruits and leaves in a single tree is lower than the inter-tree variability, we used the more robust method of measuring metabolite from fruits pooled from a tree. Although our work is based on images to predict a specific intracellular metabolite, azadirachtin, our method has potential to be applicable with a broad utility to any cell type, including three-dimensional shapes, images and architecture, towards real-time prediction of intracellular analytes.

References

1. Keasling, J. D., Synthetic biology: A global approach. *Nature* **2014,** *510* (7504), 218.
2. Minervini, M.; Abdelsamea, M. M.; Tsaftaris, S. A., Image-based plant phenotyping with incremental learning and active contours. Ecological Informatics. *Ecological Informatics* **2014,** *23*, 14.
3. Rostam, H. M.; Reynolds, P. M.; Alexander, M. R.; Gadegaard, N.; Ghaemmaghami, A. M., Image based Machine Learning for identification of macrophage subsets. *Sci Rep* **2017,** *7* (1), 3521.

4. Kovaleva, L. V.; Zakharova, E. V.; Minkina, Y., Phytohormonal control of male gametophyte growth in the pollen-pistil system. *Dokl Biochem Biophys* **2002,** *385*, 193-5.

5. Schmutterer, H., *The neem tree : Azadirachta indica A. Juss. and other meliaceous plants : sources of unique natural products for integrated pest management, medicine, industry, and other purposes*. Neem Foundation: 2002.

6. Bloice, M. D.; Stocker, C; Holzinger, A., Augmentor: An Image Augmentation Library for Machine Learning. arXiv:1708.04680 **2017**.

7. Horstmann, C. S.; Cornell, G., *Core Java*. 9 ed.; Prentice Hall: 2013; Vol. II.

8. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; 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, X., TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467v2* **2016**.

9. Schmidhuber, J., Deep learning in neural networks: an overview. *Neural Netw* **2015,** *61*, 85-117.

10. Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H., MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv:1704.04861* **2017**.

11. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z., Rethinking the Inception Architecture for Computer Vision. *arXiv:1512.00567* **2015**.

12. Smyth, N., *Android Studio 3.0 Development Essentials - Android 8 Edition*. CreateSpace Independent Publishing Platform: 2017.

13. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *Nature* **2015,** *521* (7553), 436-44.

14. Ching, T.; Himmelstein, D. S.; Beaulieu-Jones, B. K.; Kalinin, A. A.; Do, B. T.; Way, G. P.; Ferrero, E.; Agapow, P. M.; Zietz, M.; Hoffman, M. M.; Xie, W.; Rosen, G. L.; Lengerich, B. J.; Israeli, J.; Lanchantin, J.; Woloszynek, S.; Carpenter, A. E.; Shrikumar, A.; Xu, J.; Cofer, E. M.; Lavender, C. A.; Turaga, S. C.; Alexandari, A. M.; Lu, Z.; Harris, D. J.; DeCaprio, D.; Qi, Y.; Kundaje, A.; Peng, Y.; Wiley, L. K.; Segler, M. H. S.; Boca, S. M.; Swamidass, S. J.; Huang, A.; Gitter, A.; Greene, C. S., Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* **2018,** *15* (141).

15. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P., Gradient-Based Learning Applied to Document Recognition. *Proceeding of the IEEE* **1998,** *86* (11), 2278-2324.

16. Le Cun, Y.; Bengio, Y., Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks.*, Arbib, M. A., Ed. MIT: Cambridge, MA, USA, 1995; pp 255-258.

17. Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; Thrun, S., Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017,** *542* (7639), 115-118.

18. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M. C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; Kim, R.; Raman, R.; Nelson, P. C.; Mega, J. L.; Webster, D. R., Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **2016,** *316* (22), 2402-2410.

19. Lakhani, P.; Sundaram, B., Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology* **2017,** *284* (2), 574-582.

20. Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J., Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput Biol* **2017,** *13* (1), e1005324.

21. Spencer, M.; Eickholt, J.; Jianlin, C., A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction. *IEEE/ACM Trans Comput Biol Bioinform* **2015,** *12* (1), 103-12.

22. Wang, S.; Peng, J.; Ma, J.; Xu, J., Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci Rep* **2016,** *6*, 18962.

23. Quang, D.; Chen, Y.; Xie, X., DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **2015,** *31* (5), 761-3.

24. Pound, M. P.; Atkinson, J. A.; Townsend, A. J.; Wilson, M. H.; Griffiths, M.; Jackson, A. S.; Bulat, A.; Tzimiropoulos, G.; Wells, D. M.; Murchie, E. H.; Pridmore, T. P., Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. *GigaScience* **2017,** *6* (10), 1–10.

25. Chary, P., A comprehensive study on characterization of elite Neem chemotypes through mycofloral, tissue-cultural, ecomorphological and molecular analyses using azadirachtin-A as a biomarker. *Physiol Mol Biol Plants* **2011,** *17* (1), 49-64.

26. Kaushik, N.; Singh, B. J.; Tomar, U. K.; Naik, S. N.; Vir, S.; Bisla, S. S.; Sharma, K. K.; Banerjee, S. K.; Thakkar, P., Regional and habitat variability in azadirachtin content of Indian neem. *Curr Sci* **2007,** *92* (10), 7.

27. Sidhu, O. P.; Kumar, V.; Behl, H. M., Variability in Neem (*Azadirachta indica*) with respect to azadirachtin content. *J Agric Food Chem* **2003,** *51* (4), 910-5.

28. Singh, A.; Negi, M. S.; Rajagopal, J.; Bhatia, S.; Tomar, U. K.; Srivastava, P. S.; Lakshmikumaran, M., Assessment of genetic diversity in *Azadirachta indica* using AFLP markers. *Theo Appl Genet* **1999,** *99* (1-2), 8.

29. Srivastava, P.; Hazarika, R. R.; Singh, M.; Chaturvedi, R., Assessment of Age and Morphometric Parameters of Seeds on Azadirachtin Production in Neem Seed Kernels collected from various Ecotypes. *Research J Chemistry and Environment* **2010,** *14* (1), 5.

30. Veitch, G. E.; Beckmann, E.; Burke, B. J.; Boyer, A.; Maslen, S. L.; Ley, S. V., Synthesis of azadirachtin: a long but successful journey. *Angew Chem Int Ed Engl* **2007,** *46* (40), 7629-32.

31. Veitch, G. E.; Boyer, A.; Ley, S. V., The azadirachtin story. *Angew Chem Int Ed Engl* **2008,** *47* (49), 9402-29.

32. Meadows A. L.; Hawkins, K. M.; Tsegaye, Y.; Antipov, E.; Kim, Y.; Raetz, L.; Dahl, R. H.; Tai, A.; Mahatdejkul-Meadows, T.; Xu, L.; Zhao, L.,

Rewriting yeast central carbon metabolism for industrial isoprenoid production. *Nature* **2016,** *537*(7622):694.

33. Paddon, C. J.; Westfall, P. J.; Pitera, D. J.; Benjamin, K.; Fisher, K.; McPhee, D.; Leavell, M. D.; Tai, A.; Main, A.; Eng, D.; Polichuk, D. R.; Teoh, K. H.; Reed, D. W.; Treynor, T.; Lenihan, J.; Fleck, M.; Bajad, S.; Dang, G.; Dengrove, D.; Diola, D.; Dorin, G.; Ellens, K. W.; Fickes, S.; Galazzo, J.; Gaucher, S. P.; Geistlinger, T.; Henry, R.; Hepp, M.; Horning, T.; Iqbal, T.; Jiang, H.; Kizer, L.; Lieu, B.; Melis, D.; Moss, N.; Regentin, R.; Secrest, S.; Tsuruta, H.; Vazquez, R.; Westblade, L. F.; Xu, L.; Yu, M.; Zhang, Y.; Zhao, L.; Lievense, J.; Covello, P. S.; Keasling, J. D.; Reiling, K. K.; Renninger, N. S.; Newman, J. D., High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature* **2013,** *496* (7446), 528-32.

34. Ro, D. K.; Paradise, E. M.; Ouellet, M.; Fisher, K. J.; Newman, K. L.; Ndungu, J. M.; Ho, K. A.; Eachus, R. A.; Ham, T. S.; Kirby, J.; Chang, M. C.; Withers, S. T.; Shiba, Y.; Sarpong, R.; Keasling, J. D., Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **2006,** *440* (7086), 940-3.

35. eMarketer. More than a Quarter of India's Population Will Be Smartphone Users This Year. https://www.emarketer.com/content/more-than-a-quarter-of-india-s-population-will-be-smartphone-users-this-year (accessed October 18, 2018).

Table 1: Division of fruit/leaf images into ten random pairs of learning and test sets for the azaL and azaH classes.

azaL (< 0.53 %)			**azaH (> 0.65 %)**		
Set	***Set ID***	***Number of images***	***Set***	***Set ID***	***Number of images***
Learning	1	445	Learning	1	295
	2	445		2	290
	3	445		3	295
	4	445		4	290
	5	445		5	295
	6	445		6	290
	7	445		7	295
	8	445		8	290
	9	445		9	295
	10	450		10	290
Test	1	50	Test	1	30
	2	50		2	35
	3	50		3	30
	4	50		4	35
	5	50		5	30
	6	50		6	35
	7	50		7	30
	8	50		8	35
	9	50		9	30
	10	45		10	35

Table 2: Number of fruit and leaf images used for training and testing, post background removal and 7 distortions - Crop, Brightness, Gaussian/Random Blur, Rotate, Flip, Zoom and Reshape, made externally using the python module 'Augmentor' to the background removed images.

Distortion type	Image type	Training set		Test set	
		azaL	azaH	azaL	azaH
Background removal	fruit	435	297	45	30
	leaf	429	293	50	30
Crop	fruit	9,570	6,556	990	660
	leaf	9,438	6,446	1,100	660
Brightness	fruit	870	576	90	60
	leaf	858	586	100	60
Gaussian/Random blur	fruit	2,610	1,788	330	120
	leaf	2,574	2,334	300	180
Rotate	fruit	25,230	16,038	2,430	1,620
	leaf	27,032	15,823	2,700	4,320
Flip	fruit	1,740	3,802	180	300
	leaf	1,726	2,888	200	320
Zoom	fruit	870	1,192	180	120
	leaf	1,716	1,172	200	120
Resize	fruit	6,960	5,364	810	540
	leaf	7,722	5,274	900	540

Table 3: Final fruit (orange) and leaf (green) models.

Image type	Architecture	Image resolution	Width multiplier	No. of training steps	azaL classification sensitivity (%)	azaH classification sensitivity (%)	Average classification sensitivity (%)
fruit	InceptionV3	-na-	-na-	500	82.81	81.56	82.18
fruit	InceptionV3	-na-	-na-	2500	80.87	84.89	82.88
leaf	MobileNet	224	0.25	1000	70.00	80.00	75.00
leaf	MobileNet	224	0.25	4500	84.00	73.33	78.67
leaf	MobileNet	192	0.75	3000	76.00	70.00	73.00
leaf	MobileNet	192	1.00	5000	76.00	70.00	73.00
leaf	MobileNet	128	0.5	2500	92.00	73.33	82.67
leaf	MobileNet	128	0.75	500	86.00	73.33	79.67

Table 4: Independent validation using independently collcted fruit images.

Measured class	Predicted class
azaH	azaL
azaL	azaL
azaH	azaH
azaH	azaH
azaH	azaH
azaH	azaL
azaH	azaL
azaH	azaL
azaH	azaH
azaH	azaL
azaH	azaL
azaH	azaL
azaH	azaH
azaL	azaL
azaL	azaL
azaL	azaL
azaH	azaH
azaH	azaH
azaL	azaL
azaH	azaH
azaL	azaL
azaL	azaL
azaH	azaH
Classification sensitivity (%)	69.57