Parametric Model for Flora Detection in Middle Himalayas

Aviral Sharma, Department of Informatics, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India*

Saumya Nigam, University of Petroleum and Energy Studies, India

ABSTRACT

Plant detection forms an integral part of the life of the forest guards, researchers, and students in the field of botany and for common people also who are curious about knowing a plant. But detecting plants suffer a major drawback that the true identifier is only the flower, and in certain species, flowering occurs at major time period gaps spanning from few months to over 100 years (in certain types of bamboos). Machine learning-based systems could be used in developing models where the experience of researchers in the field of plant sciences can be incorporated into the model. In this paper, the authors present a machine learning-based approach based upon other quantifiable parameters for the detection of the plant presented. The system takes plant parameters as the inputs and will detect the plant family as the output.

KEYWORDS

Machine Learning, Plant Recognition

INTRODUCTION

The present era is been led by artificial intelligence and machine learning which has enabled us to use the existing technologies in virtually all walks of life and solve real-world problems. Machine learning is the sub-domain of artificial intelligence that allows machines to replicate the behavior of learning. As humans perform learning from observations, machines can also do the same when supplied suitable examples for the same.

A large increase has been observed in the potential applications that could use the various machine learning algorithms where the machine can be trained to perform various tasks and give unexpected results(Singh et al., 2016). All the industries have been accumulating data since the time of their inception and the pace of data generation has gone up in the modern days and hence the requirement for a fast-paced analysis tool. Applications of machine learning have been expanding and it is now used from oil and gas to transportation, financial services, and health care to government organizations. The ultimate goal of building these applications is to utilize the capabilities of machine learning for process automation to increase productivity and reduce human efforts to facilitate everyday tasks.

Plant species recognition is one of the complex tasks to be performed in everyday life which involves several challenges given the diverse biodiversity present. The complexity increment exponentially if the given job is to be performed for any bio-diversity hotspots especially the Himalayan region which has been bestowed with an extremely varied gene pool and species, stretched over a geographic area of four million square kilometers. Artificial intelligence and machine learning thus find a potential application in this area. The recent times have observed a significant expansion and

DOI: 10.4018/IJDSST.286698
*Corresponding Author

increment of various approaches that could be useful for plant species recognition, detection of weeds in plants, etc. The research on these topics has the final goal of fully automating the agricultural process beginning from detecting the species itself. While the final goal is still to be achieved, significant advancements in the area have been made. The proposed algorithm needs to be precise and should increase productivity with the least margin of error possible.

PLANT PARAMETERS

Eight parameters were selected for plant detection based upon the inputs from the concerned subject matter expert. These are as follows:

HABITAT: Habitat is the home of any particular type of species. It the place where a particular type of species can be found. As we can observe from numerous maple species that they always have a shrubby habit and almost always form bushes or hedges but never a tree. Certain alpine plants have been chosen for cultivation because of their dwarf habit. A few plants created wood as an instrument to neutralize the dangerous impacts of wind, ice, mechanical harm, and flame. Erect and thick development propensities advanced to oppose wind impacts and other mechanical harms. Plants without wood adjusted prostate, tangle framing, spreading, crawling, or climbing propensities. As creatures connected with plants and before, both developed at the same time. Different plants created prostrate and tangle shaping propensities so as to persevere through exceptional brushing and trampling, or erect and tall development structures to avoid perusing and touching.

HEIGHT: The height of a plant makes one of the distinguishing features especially between varieties. It the shortest distance from ground level to the topmost photosynthetic tissues of the plant. This is usually expressed in meters. While measuring the height of a plant, for any species, the under consideration entity is a typical mature entity of the said species in a particular habitat. Maximum height is associated with numerous growth factors like the position of the plant as compared to the vertical light gradient, reproductive size, competitive vigor, whole-plant fecundity, lifespan and allied between two different events of disturbance like a wildfire. Plant propensity is likewise called vegetation structure. It is a procedure that all plants experience that is a type of transformative adaptation. It misrepresents the plant's capacity to change, endure, and eventually effectively recreate regardless of consistently evolving conditions. The plant propensity provides some insight into how the plant has advanced to withstand even emotional climatic swings, utilization by herbivores, loss of daylight, and increment/decline in water. The term likewise alludes to how well a plant, at last, recreates and colonizes in worldwide locales. For instance, succulents can endure desert-like conditions because of its plant propensity. Succulent plants built up their interesting appearance through developmental survival in brutal conditions, i.e., their advancing plant propensity. Succulents can exploit brisk, inconsistent precipitation cycles because of their capacity to hold and store water – these characteristics are for the most part because of plant propensities.

BARK TYPE: Bark is the outermost layer of the trunk or roots of the tree. It is comparative from multiple points of view to human or animal skin. Underneath the bark of the tree, cambium is found. Each developing season adds a layer of Xylem cells around the cambium which encompasses the cambium. Xylem cells act as the mode of transport for the movement of minerals from trees to root. It comprises dead cells and constitutes a significant part of the tree. This layer is inside the cambium. The bark is essentially made up of the few layers that are outside the cambium. The external edge of the cambium delivers another set of cells that make phloem, which delivers sugars from the leaves to the rest of the tree. Outside phloem, most of the trees consist of a layer known as the plug cambium, which creates the stopper – the intense external layer of the tree. This external layer is all that we, for the most part, observe the bark. The external stopper shields the tree from the components - from burning by the sun or drying by the wind. It additionally averts parasites and the numerous creepy crawlies and warm-blooded creatures that would somehow or another exploit the sugar-rich sap or the wood that it encompasses. The outermost layer of roots and stems in plants makes up the bark. It's mostly found in woody plants which include but are not limited to trees, woody vines, and shrubs.

BARK COLOR: Considerably more than the schematics of its layers in a green bean science book, the bark is the covering, the existence support, and the essence of a tree, permanently scratched with its long periods of abundance and hardship. Every species has its very own extraordinary bark design, frequently sufficiently unmistakable to fill in as a method for distinguishing proof. It tends to be red, green, dim, white, orange, or striped; prickly, smooth, unpleasant, or profoundly wrinkled; or it can strip away to make a diverse woven artwork. Seen close up and in detachment from the remainder of the plant, bark welcomes correlation with theoretical workmanship. As a plan component in the greenery enclosure, the bark is the last boondocks, intriguing in the winter positively, yet in addition, a reward of surface and example all year that improves foliage and blossoms. Few trees are light gray; whereas some others are dark gray. Few trees have white trunks with streaks of black like paper birch. The Palo Verde has green bark. The bark of other trees, like cedars and redwoods, do look brown, with shades of gray

LEAF TYPE: The leaf is one of the most important factors in the identification of the plant. Observations in the type of leaves have to be taken from the petiole which can actually tell the type of leaf. Starting from the petiole, the leaf is of two different types:

Simple leave - A simple leaf consists of a single, entire, or dissected lamina. The simple leaf has got all the structures of a typical leaf.

Compound leaf - A leaf is said to be compound when the leaf blade is dissected into a number of small segments called leaflets.

LEAF LENGTH: Once the type of leaf is known, the other parameter of importance is the length and the width of the leaf as there is variation between these as well. While conducting this work, the length of the leaf (L) and width of the leaf (W) of the leaf sharp edge was estimated with a millimeter ruler.

LEAF SHAPE: The state of a leaf can likewise give signs while recognizing broadleaf tree species. Basic shapes incorporate applaud (egg-formed), lanceolate (long and restricted), deltoid (triangular), orbicular (round), and cordate (heart molded). Tree leaves depict the long term evolution of a particular species to the exposures in any particular land area. These are used to infer the ecosystems limiting factors that have influenced the final form and shape of a tree's leaves.

INFLORESCENCE: An inflorescence is a gathering or group of blossoms masterminded on a stem that is made out of a fundamental branch or a muddled course of action of branches. Morphologically, it is the altered piece of the shoot of seed plants where blooms are shaped. The alterations can include the length and the idea of the internodes and the phyllotaxis, just as varieties in the extents, compressions, swellings, adnations, connotations, and decrease of primary and optional tomahawks. Inflorescence can likewise be characterized as the conceptive segment of a plant that bears a bunch of roses in a particular example. The stem holding the entire inflorescence is known as a peduncle and the real pivot (mistakenly alluded to as the primary stem) holding the blossoms or more branches inside the inflorescence is known as the rachis. The stalk of every single bloom is known as a pedicel. Abloom that isn't a piece of an inflorescence is known as a single blossom and its stalk is likewise alluded to as a peduncle. Any blossom in an inflorescence might be alluded to as a floret, particularly when the individual blooms are especially little and borne in a tight group, for example, in a pseudanthium. The fruiting phase of an inflorescence is known as an infructescence

DATA COLLECTION

The Himalayas are covered with a different array of flora and fauna stretched over an area of four million square kilometers. Our data collection for the project “Parametric model for flora detection in the middle Himalayas” has been done in the Chakrata region in Dehradun district, Uttarakhand. It is between the Tons and Yamuna rivers, at an elevation of 7000–7250 feet, 98 km from the state capital, Dehradun. Data collection included gathering samples from various plants of Chakrata region including stems, leaf, images of plants and their barks, etc. and all the samples were taken to **Forest Research Institute (FRI),** Dehradun. The data were distinguished from each other according to their species and unrelated data was discarded. a further detailed study of the selected plants and information of their properties was done, parameters were created such as leaf type, leaf shape, inflorescence, bark color, bark type, the height of the plant, etc. and thus, useful data for the project was extracted from all the samples(Crisci et al., 2012).

After a detailed analysis of the data and verification, the **final dataset** was created according to various parameters that distinguish one plant from another.

AI AND ML IN PLANT SCIENCES

Engineers have been using machine learning techniques to train the models in order to make smart decisions in the future. Machine Learning techniques have their application in various places like Image recognition, speech recognition, robotics, financial services, etc. Machine learning techniques have been used in the field of plant sciences as well. Petre Lameski (2017) used machine learning and Image Processing for plant species recognition. In this research, Petre’s goal was the detection and segmentation of unwanted weed plants. For this purpose, Petre recorded several datasets from seedling plantations in the Republic of Macedonia using common, commercially available cameras. Petre generated two datasets, one with tobacco seedling images and the other with spinach, carrot, and salad images, and applied machine learning algorithms on this data set. As a result, this approach was the first step of the design and implementation of a low cost, machine learning-based sensor system, that could be used for weed control automation. Khwaja Osama, Bhartendu Nath Mishra, and Pallavi Somvanshi have expressed their idea of using Machine Learning Techniques(2015) in Plant Biology in the book Phenomics: Technologies and Applications in Plant and Agriculture where they discussed the need for analyzing the immense amount of biological data and it’s integrating with information of plant biology like crop improvement, biochemical engineering, etc. They also discussed a few machine learning approaches and their applications in plant biology. Oluwafemi Tairu (2018) built a plant disease detection model using the Convolutional Neural Network. For this Tairu loaded the data set with images of diseased plants and applied machine learning techniques to it. As a result, this model had an accuracy score of 96.77% and later this model was deployed for use as a Mobile App through an API. A very similar type of application was developed at Google by Shaza Mehdi and Nile Ravendell (2018) using TensorFlow which is an open-source machine learning library that helped farmers in identifying the diseased plants. In this app, identification of a diseased plant could be done by waving their phone in front of the pant. The application can also provide options for the best ways to manage it. A similar Mobile-Based Deep Learning Model for Cassava Disease Diagnosis(2019) has been developed by Amanda Ramcharan, Peter McCloskey, Kelsee Baranowski, Neema Mbilinyi, Latifa Mrisho, Mathias Ndalahwa, James Legg, and David P. Hughes. In this study, the performance of a CNN model has been evaluated where the model is deployed in real-time on a mobile device in order to detect symptoms of cassava pests and disease. A single-shot detector model has been used in order to optimize CNN architecture for mobile devices and assessing its performance to detect various symptoms of many disease classes. For comparing the test dataset results in real-world images and videos, the decrease in its F-1 score was used.

In the next section, we will be talking about our parametric model, which incorporates the use of machine learning algorithms to make decisions based on the parameters provided to it. The data set of the parametric model has active backing from the researchers and experts from the respective domain. The model takes plant parameters as inputs and detects the plant family as output.

DESIGN OF PARAMETRIC MODEL

Machine learning deals with classification, the research work is used to classify healthy and unhealthy plants. Our work is based on morphological features of the plant leaf. The techniques represented in this paper are

1. Decision Tree
2. K- Nearest Neighbour
3. Naive Bayes
4. Support Vector Machines
5. Multi-Level Perceptron

DECISION TREE

A decision support tool comprising of a tree-like structure of all the decisions and possible consequences is the decision tree. Its structure is similar to a flowchart where each node corresponds to a test on the respective attribute and every branch constitutes the outcome of that test(Freund & Mason, 1999). The source code is split into various subsets based on the value attribute test. The same process is applied to every subset in a repeated manner by performing recursive partitioning. This tree takes an item or circumstance depicted by a set of properties and yield as Yes/No.

Entropy and Info Gain in a decision tree:-

Entropy in a decision tree helps in controlling how the data is been split into a decision tree. The boundaries of a decision tree are affected by its entropy.

$$Entropy = -\sum p\left(x\right) logp\left(x\right) . ... \qquad (1)$$

Information gain is the measure of the information that a feature is capable of providing about a class. This is the main key that is needed in building an algorithm in order to construct a decision tree. A decision tree will always try to maximize its value in the algorithm

K-NEAREST NEIGHBOUR

K-Nearest Neighbors is one of the most fundamental yet very essential components that is used in calculations in Machine Learning. It has a unique place among various machine learning algorithm which finds application in most of the industry problems. It is a type of supervised algorithm that is used in classification as well as regression problems. It is defined by two properties:

- Lazy learning algorithm
- Non-parametric learning algorithm

The training phase in a KNN is not specialized and all the data is used during the classification period(Zhang & Zhou, 2005). The algorithm is non-parametric and does not make any assumptions about the given data. It uses the similarity feature in order to predict new data points on the basis of their close matching with the data points of the training set. Searching is performed on the entire dataset to make predictions by looking for k most similar neighbors. Euclidean distance is one of the most common methods used for calculating the distance between k-most similar neighbors and also summarizes the output variables for them.

$$Elucidean_{distance\left(x,x_i\right)} = sqrt\left(sum\left(x_i - x_{ij}\right)^2\right). \dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (2)$$

One disadvantage of KNN is that it can get time-consuming with the increasing size of the dataset as it calculates the distance between every data point.

NAIVE BAYES

It is a type of supervised machine learning algorithm. It does not constitute a single algorithm, rather a family of classification algorithm where every pair of features is considered to be independent of each other. It is based on the Bayes theorem, which gives the relationship between a given class variable y and the dependent vector given through x_1 to x_n:

$$P\left(y|x_1,\dots\dots\dots,x_n\right) = \frac{P\left(y\right)P\left(x_1,\dots\dots\dots x_n|y\right)}{P\left(x_1,\dots,x_n\right)} \,. \dots\dots\dots\dots\dots \quad (3)$$

By using the conditional independence assumption:

$$P\left(x_i|y,x_1,\dots\dots,x_{i-1},x_{i+1},\dots x_n\right) = P(x_i \mid y) \,. \dots\dots\dots\dots\dots\dots\dots\dots\dots.. \quad (4)$$

Relation simplifies to:

$$P(y|x_1,\dots,x_n) = \frac{(P\left(y\right)\Pi_{i=1}^{n} P(x_i \mid y))}{P\left(x_1,\dots,x_n\right)} \,. \dots\dots\dots\dots\dots\dots\dots.. \quad (5)$$

There are various naive Bayes classifiers which differ in terms of their assumptions. Some of the classifiers include:

- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Bernoulli Naive Bayes

SUPPORT VECTOR MACHINES

Support Vector Machines come under the supervised machine learning models that are often used for classification and regression analysis. They use the concept of decision planes which are used to define boundaries between the various decisions. This type of plane is used to separate various sets of objects belonging to different classes(Hsu et al., 2003). One such example is shown in the

Figure 1. Hyperplane

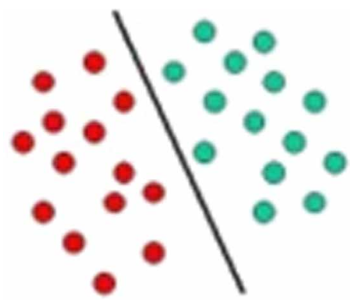


illustration below. In the following example, the objects belong to two different class memberships which are GREEN or RED. The separating line represents one such boundary where the objects on its right are GREEN and those to the left belong to the RED class. Any object belonging to the right of the boundary is classified with a label of GREEN class and others on the left are given the label of RED class as depicted in Figure 1.

The objective of the above algorithm is to define a hyperplane in N-Dimensional space, where N is the number of features so that each data point can be classified distinctly. In order to separate two classes, many hyperplanes can be selected. However, the goal is to select the hyperplane that can maximize the distance between the data points of both the classes i.e. maximize the margin.

Data points closer to the hyperplane are termed as support vectors and they tend to influence its orientation. The loss function is used to maximize the margin.

$$c\left(x,y,f\left(x\right)\right)=\begin{cases}0, & if\, y*f\left(x\right)\geq 1\\ 1-y*f\left(x\right), & else\end{cases}\ldots\ldots.. \tag{6}$$

We calculate the loss value if the predicted value and the actual value does not have the same sign i.e. cost is not zero.

MULTILAYER PERCEPTRON

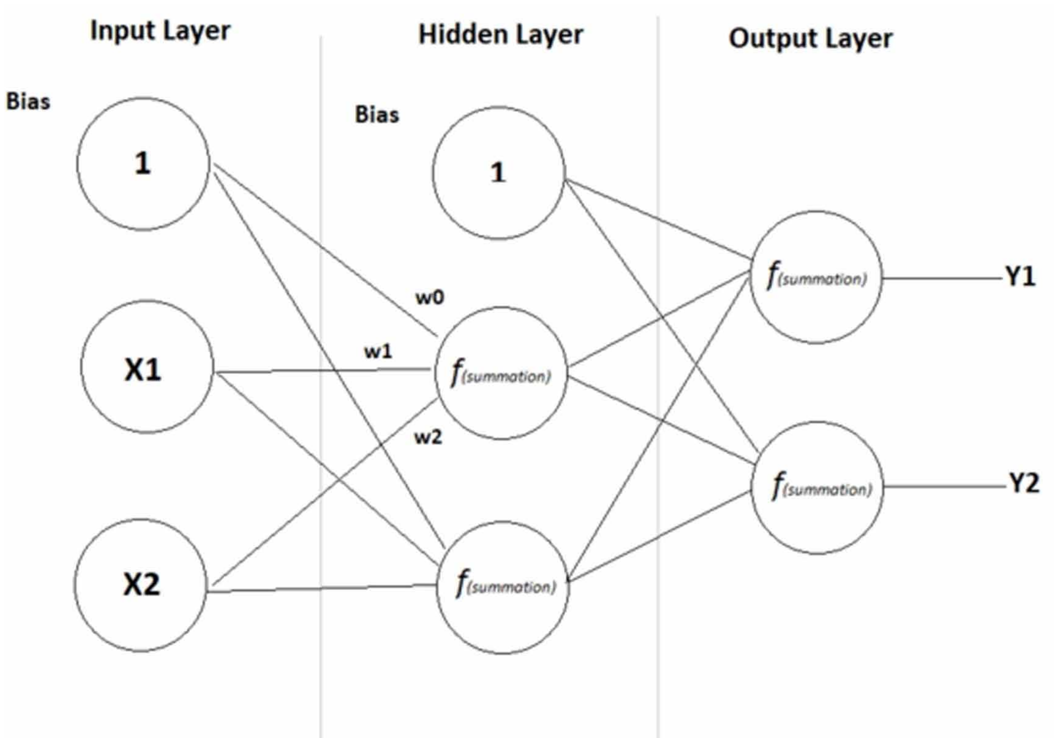
It is a type of deep artificial neural network that consists of more than one perceptron. It has one input layer, one output layer for receiving signals making predictions from those signals. Another layer is used in between called the hidden layer which is used in performing many complex tasks(Karlik & Vehbi Olgac, 2011). It is a feedforward network because the output from one layer is used as an input for the next layer. The number of hidden layers can be increased in accordance with the complexity of the task.

There are 3 steps used in training:

- Forward pass
- Calculate error or loss
- Backward pass

The bias is applied to all the weights in the first step and the output is calculated which is fed to the next layer. Various loss functions are then applied in order to calculate the deviation of the predicted output from the expected output i.e. loss. Backpropagation is then performed in order to reduce it.

Figure 2. Artificial Neural Network



RESULT

The dataset used in this approach made use of 8 plant parameters to perform the differentiation between the various species. To each of these plants in the dataset, a testing set was used to test the accuracy of each of the specified models. A number of incorrect recognition is listed in the last column of each of these tables. The respective plant species with the correct number of recognitions for the respective model is also defined. During the course of the study, we have seen that KNN provided the highest accuracy and MLP provided the lowest accuracy.

The performance of each of the proposed approach is evaluated using the testing dataset and the real dataset based on the accuracy and efficiency of each of the classifiers.

CONCLUSION

A new approach to plant identification has been proposed in this paper. This work is especially useful in cases when the flower of the plant is not available and a can help a laymen recognize the plant it is also to important mention here that the prime identification of any plant depends on its flower since trees have a longer life span and very short flowering period a parameterized model can help plants to be identified by machines with high amounts of certainty.

Table 1. SVC

Scientific Name	Common Name	Training Sample	No. of Incorrect Recognition
S.Robusta	Sal Tree	10	9
F.Ramontchi	Batoko Plum	10	0
C.Fistula	Golden Shower	10	0
M.Philippinensis	Kumkum Tree	10	0
A.Cordiflia	Baby Son Rose	10	0
C.Infortunatum	Hill Glory	10	0
A.Vasica	Adhatoda	10	0
L.Camara	Tickberry	10	0
F.Religiosa	Peepal Tree	10	0
Murraya	Maraaya	10	0

Table 2. MLP

Scientific Name	Common Name	Training Sample	No. of Incorrect Recognition
S.Robusta	Sal Tree	10	10
F.Ramontchi	Batoko Plum	10	10
C.Fistula	Golden Shower	10	10
M.Philippinensis	Kumkum Tree	10	10
A.Cordifolia	Baby Son Rose	10	10
C.Infortunatum	Hill Glory	10	10
A.Vasica	Adhatoda	10	0
L.Camara	Tickberry	10	10
F.Religiosa	Peepal Tree	10	10
Murraya	Marraya	10	10

Table 3. KNN

Scientific Name	Common Name	Training Sample	No. of Incorrect Recognition
S.Robusta	Sal Tree	10	1
F.Ramontchi	Batoko Plum	10	0
C.Fistula	Golden Shower	10	0
M.Philippinensis	Kumkum Tree	10	0
A.Cordifolia	Baby Son Rose	10	0
C.Infortunatum	Hill Glory	10	0
A.Vasica	Adhatoda	10	0
L.Camara	Tickberry	10	0
F.Religiosa	Peepal Tree	10	0
Murraya	Marraya	10	0

Table 4. Decision Tree

Scientific Name	Common Name	Training Sample	No. of Incorrect Recognition
S.Robusta	Sal Tree	10	1
F.Ramontchi	Batoko Plum	10	0
C.Fistula	Golden Shower	10	0
M.Philippinensis	Kumkum Tree	10	1
A.Cordifolia	Baby Son Rose	10	0
C.Infortunatum	Hill Glory	10	0
A.Vasica	Adhatoda	10	0
L.Camara	Tickberry	10	0
F.Religiosa	Peepal Tree	10	0
Murraya	Marraya	10	0

REFERENCES

Aggarwal, C. (2018). *Neural networks and deep learning*. Springer.

Crisci, C., Ghattas, B., & Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling*, *240*, 113–122. doi:10.1016/j.ecolmodel.2012.03.001

Freund, Y., & Mason, L. (1999). *The alternating decision tree learning algorithm.* ICML.

Hsu, Chang, & Lin. (2003). *A practical guide to support vector classification*. Academic Press.

Karlik, B., & Vehbi Olgac, A. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, *1*(4), 111–122.

Nielsen, M. A. (2015). Neural networks and deep learning. San Francisco, CA: Determination Press.

Osama, K., Mishra, B. N., & Somvanshi, P. (2015). Machine learning techniques in plant biology. In *PlantOmics: The Omics of plant science* (pp. 731–754). Springer.

Ramcharan, A., McCloskey, P., Baranowski, K., Mbilinyi, N., Mrisho, L., Ndalahwa, M., Legg, J., & Hughes, D. P. (2019). A mobile-based deep learning model for cassava disease diagnosis. *Frontiers in Plant Science*, *10*, 272. doi:10.3389/fpls.2019.00272 PMID:30949185

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom),* 1310-1315.

Zhang, M.-L., & Zhou, Z.-H. (2005). *A k-nearest neighbor based algorithm for multi-label classification. In 2005 IEEE international conference on granular computing* (Vol. 2). IEEE.