Determining the Effectiveness of Soil Treatment on Plant Stress using Smart-phone Cameras

Anurag Panwar*, Mariam Al-Lami†, Pratool Bharti*, Sriram Chellappan*, Joel Burken†
*Department of Computer Science and Engineering
University of South Florida, Tampa, Florida 33620, USA.
{anuragpanwar, pratool}@mail.usf.edu, sriramc@usf.edu
†Department of Civil, Architectural and Environmental Engineering
Missouri University of Science and Technology, Rolla, Missouri 65409, USA.
{mkan87, burken}@mst.edu

***Abstract*—Plants are vital to the health of our biosphere, and effectively sustaining their growth is fundamental to the existence of life on this planet. A critical aspect, which decides the sustainability of plant growth is the quality of soil. All other things being fixed, the quality of soil greatly impacts the plant stress, which in turn impacts overall health. Although plant stress manifests in many ways, one of the clearest indicators are colors of the leaves. In this paper, we conducted an experimental study in a greenhouse for detecting plant stress caused by nutrient deficiencies in soil using smart-phone cameras, coupled with image processing and machine learning algorithms. The greenhouse experiment was conducted by growing two plant species; willows (*Salix Pentandra*) and poplars (*Populus deltoides x nigra, DN34*), in two treatments. These treatments included: unamended tailings (collected from a lead mine tailings pond and characterized by nutrient deficiency), and biosolids amended tailings. Biosolids are very rich in nutrients and were added to the tailings in one of the two treatments to supply plants with nutrients. Subsequently, we captured various images of plant leaves grown in both soils. Each image taken was pre-processed via filtration to remove associated noise, and was segmented into pixels to facilitate scalability of analysis. Subsequently, we designed random forests based algorithms to detect the stress of leaves as indicated by their coloring. In a dataset consisting of 34 leaves, our technique yields classifications with a high degree of prediction, recall and F1 score. Our work in this paper, while restricted to two types of plants and soils, can be generalized. We see applications in the emerging area of urban farming in terms of empowering citizens with tools and technologies for enhancing quality of farming practices.**

***Keywords*—Plant stress, Participatory sensing, Smart-phones, Image processing, Machine learning algorithms**

I. Introduction

Vegetation offers a broad benefit to society, primarily as our counter parts in the global carbon and water cycle by producing all the oxygen we need, and also serving as a critical source of our food. Plants are also vital in connecting our terrestrial world with the atmospheric water cycle. A number of studies are being conducted today to enable healthier growth of plants in our biosphere. A critical factor that decides the overall health and sustainability of plants is the soil in which they grow. Numerous studies have shown that plants grown in superior (i.e., nutrient rich) soils are much more healthy, resistive to disease and live longer than plants that grow in nutrition deficient soils [1][2][3][4][5], and as such understanding the overall health of plants in relation to the soil in which they grow is important.

A critical component of plant health is decided by plant stress responses, which are characterized by a suite of molecular and cellular processes that are triggered by the plant in some form of stress. Stresses can be abiotic, such as drought or excess light, or biotic, such as herbivores or pathogens. Studies have shown that, while plants do indicate their stress via variation of leaf in chemical content [6] and molecular changes [7] [8], the colors of leaves are also clear indicators of plant stress [9][10][11]. In particular, deficiency of nutrients such as nitrogen, phosphorous, potassium, calcium, and magnesium in soils results in changes in coloring patterns of leaves. Healthier/Stress free leaves are typically greener in color, while increased yellowing indicates progressively un-healthier leaves [9][10][11]. Unfortunately, existing techniques to detect plant health and stress incur expensive infrastructure, and/ or significant manual effort.

In this paper, we are broadly motivated to leverage off-the-shelf smart-phones, and their in-built cameras to assess plant stress based on the soil quality, (both nutrient deficient and nutrient rich). In our experiments conducted in a greenhouse facility, we grew laurel-leaf willows (scientifically known as *Salix Pentandra*) and poplars (*Populus deltoides x nigra, DN34*) in two treatments - unamended tailings and tailings amended with biosolids. Tailings are waste product associated with mining activities of extracting economic minerals from the ore. These tailings usually lack some of the essential nutrients and have very poor soil structure. To improve soil quality, we considered biosolids as an amendment to be added to some of these tailings soils to improve nutrition. Subsequently, a total of 34 leaves were imaged using the built-in camera of a Samsung GALAXY S4 phone, with the goal being to identify plants stress based on the coloring of the leaves. Within this scope, our contributions are the following.

a. Image Preprocessing Techniques: Due to limited processing power of smartphones, image resizing is important. In our analysis, we determined that resizing images from 4182×2322 pixels (that is typical in the phone) to 413×233 pixels greatly saved on computational overhead without compromising accuracy significantly. Then, filtering techniques like mean, median and sharpening filters were used to remove unwanted noise. Filtering takes care of variation of

brightness and color information up to certain degree. Then image segmentation was utilized to get region of interest which in our case are leaves, that makes analysis easier and faster by reducing the number of pixels.

b. Learning Algorithms based on Random Forests: We designed a classification algorithm based on Random Forests [12][13][14] for classifying plant stress. Random forests is an ensemble supervised machine learning technique which uses decision tree as base classifier. For every single tree, feature/split-points are randomly selected from set of features based on information gain. It uses decision tree which is very fast in predicting results which is an important requirement when we attempt classification problems involving big data sets like image data.

c. Evaluation : We evaluated our algorithm on 34 leaves taken from plants grown in both soils. Of these, images from 15 were used for training and 19 for testing. Based on cross validation analysis, we demonstrated that our algorithm achieves a high degree of precision, recall and F-1 score.

We point out that our work in this paper is restricted to only two types of plants and soils. However, with more experiments, we can generalize our technologies much more for wider applicability. Another area where we believe our work will be impactful is *urban farming*, that is receiving a lot of attention today to enable greener societies with numerous benefits. For instance, studies show that every $1.00 invested in community plots can yield upto $6.00 worth of vegetables. Food produced in urban farms is fresher and retains nutritional value better, compared to food that travels. Urban farms save energy and water in multiple ways, and have shown correlations with economic prosperity and happier societies. We believe that our work in this paper provides foundations for innovative tools and technologies for urban farmers of tomorrow, and exploring this application in more detail is part of our on-going work.

II. Related Work

In this section, we highlight important work on using image processing techniques to assess plant stress. Lindow and Webb [9] proposed thresholding based methods on images of plants captured using analog video cameras under a red light illumination to detect necrotic areas of the plant. In [10], thresholding and segmentation methods were used to quantify the severity of coffee leaf rust on plants with images taken from both black & white, and color charge coupled device (CCD) cameras. Similar techniques using CCD cameras are also proposed by Story *et al.* [15] to detect calcium deficiency in lettuce . Carter *et al.* [16] also used CCD cameras for early detection of plant stress. The experiments involved collecting images and physiological measurements of stress due to factors like rainfall, exposure to herbicide etc. For digital imaging, leaf fluorescence and reflectance value at certain wavelengths were used for classification. All these techniques unfortunately require specialized equipment that is cost prohibitive for broader applications and requires considerable expertise, and are also time and resource intensive.

A number of techniques have also been proposed to use commercially available digital cameras to assess plant stress. Bock *et al.* [17] shown that images of leaves when captured and processed using state-of-the-art photoshop techniques do reveal insights on plant stress. The issue though is the amount of manual inspection involved. Abdullah *et al.* [11] proposed a method which discriminates a given disease (corynespora) from other pathologies that affect rubber tree leaves. In the proposed approach Principal Component Analysis is applied directly to the RGB values of the pixels of a low resolution (15×15 pixels) images of the leaves, which were obtained by FinePix 6900 Zoom (Fuji Film) digital camera. The first two principal components are then fed to a Multilayer Perceptron (MLP) Neural Network with one hidden layer, whose output reveals if the sample is infected by the disease of interest or not. The resolutions of images used here were quite low, that leads to reduced accuracy of classification, and the proposed neural network approach is computationally very expensive. Another technique proposed by Sena *et al* [18] discriminates between maize plants affected by fall armyworm from healthy ones using images captured from digital cameras. Their approach pre-processes the images and partitions them into 12 blocks, and counting the number of connected diseased regions are used as indicators of plant stress. This approach requires significant training to determine thresholds and also requires manual inspection, which limits its practicality.

Comparing our Work w.r.t. Related Work: At the outset, the core novelty of our work lies in using off-the-shelf smartphone (i.e., Samsung GALAXY S4) cameras to detect stress in plant leaves. As such, the generating more images (for superior training and classification) is much more easy now. However, there are some challenges with smart-phone cameras (and possibly un-trained citizens using these cameras), which we need to resolve. The first challenge comes from the limited processing capability of phones, for which appropriate image compression needs to be employed, without compromising accuracy of classification. Our analysis revealed that a tenfold reduction in images provides a high degree of classification accuracy, while greatly minimizing overhead. The second challenge comes from noise associated with image, for which we use mean, median and sharpening filters, which have relatively low complexity, and provide significant reductions in overall noise.

Also, with more images, there is associated problem of scalability of algorithmic techniques to process these images. Image segmentation was employed to get regions of interest which in our case are leaves, that makes analysis easier and faster by reducing the number of pixels that need to be processed. Finally, any algorithm designed must be accurate and robust in classification of plant stress. While traditional techniques like Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) can be applied, KNN is constrained by the size of dataset. KNN is a deterministic classifier with no stochastic property which results in low accuracy and overfitting problem. KNN uses whole dataset for training which creates the problem of overfitting. In KNN, frequent class tends to dominate the prediction of the new data, because they tend to be common among the k nearest neighbors due to their large number. KNN also suffers from the problem of k parameters. Selecting the correct value of fixed k is very difficult task. State-of-the-art SVM also have some problems. SVM is very slow for large dataset and does not provide immunity to outliers. We employ a Random Forests Based approach which uses bagging technique. It provides immunity to outliers and harnesses the power of randomness property in

case of large datasets for superior accuracy. Instead of selecting whole dataset for training, it divides the data into several mini-batches and uses them for training separate decision tress and then uses voting to select the class with most votes.

The overall technique we propose in this paper (images capturing, filtering, segmentation, and classification) can be easily encoded as a simple to use smart-phone app for wide spread adoption (which is part of on-going work), and it is an added feature of our work in this paper.

III. Experimental Setup and Data Collection

Our experiments were conducted for three months in Spring 2015 at a greenhouse facility situated in Missouri University of Science and Technology. Two soils were identified for the study first. The first soil was a tailing soil collected from a lead mine tailings pond. These tailings are waste product associated with mining and smelting activities of extracting economic minerals from the ore, as a result of which the soil is nutrition deficient for sustaining healthy plant growth. More specifically, the tailings lack organic matter and plant nutrients, and characterized by slightly high pH. Availability of most minerals essential for plant growth is strongly affected by soil pH. As soil pH increases, bioavailability of nutrients such as P, K, Fe, Mn, Zn, Cu decreases. [19][20].

The second type of soil used in this study was the tailing soil treated with biosolids. The biosolids used in the experiment were prepared by drying sewage sludge collected from a water treatment plant. The biosolids were rich in organic matter and have a full range of plant nutrients, necessary for sustaining plant growth. The biosolids also reduce the presence of lead and nickel that are harmful for plant growth[1].

Two types of tolerant plant species were carefully identified for the study[2]. These two plant species were laurel-leaf willow (scientifically known as *Salix Pentandra*) and poplar (*Populus deltoides x nigra, DN34*). They were both grown separately for a period of three months in both the tailing soil and the tailing soil treated with biosolids. Tailings were air dried and crumpled to pass through a 2 mm sieve. Tailings were placed into 2 liter pots (1350g per pot). To improve the soil nutrition, biosolids were added at a rate equivalent to 40 dry ton/acre and mixed thoroughly with the tailings. Pots were prepared in triplicates for each treatment resulting in a total of 12 pots. The plant cuttings were planted in triplicate in each pot resulting in a total of 18 cuttings for each plant species. Images were taken at 3 months growth period.

The image data for our experiments contains leaf samples from both plants grown in both soils in the greenhouse. A Samsung GALAXY S4 smartphone was used for capturing images collected in regular daylight. A total of 34 images of leaves were collected from the smartphone spread across both plants grown in both soils. The phone has following camera configuration, shown in **Table I**.

[1]Please refer to Appendix for a list of important minerals present in the original tailing soil, and biosolids used in the study.

[2]Tolerance is a term that ecologists use to indicate a tree's capacity to develop and grow in the shade of, and in competition with, other trees.

Algorithm 1 Algorithm for pre-processing and feature selection of image data

Pre-processing steps :

Resizing Raw input images taken from smartphone: $I_i^{M,N}$
Re-sized image: $I_i^{m,n}$ where M and N are the pixels in x and y direction for the image before re-sizing. m and n after resizing. $M > m$ and $N > n$

Filtering Input: $I_i^{m,n}$+$\delta_i^{m,n}$
Output: $I_i^{m,n}$
$I_i^{m,n}$ depicts image with minimal amount of noise and $\delta_i^{m,n}$ represent noise associated with image i

Segmentation Input: $I_i^{m,n}$
Output: Segmented Image

Feature selection Input: Pre-processed images $\vec{I}$ = $\{\vec{I}_i^{m,n}\}$
Output: Features selection $f_1, f_2...., fn$ where n is the total number of relevant features
$\vec{B}_i$ = *FeatureSelection*($\vec{I}_i^{m,n}$) $\forall i \in (1, N)$ where N is the total number of images;
$\vec{B}_i$ = $f_1, f_2...., fn$

TABLE I: Camera specification of Samsung GALAXY Smartphone used for experiments

Camera Specification	Value
Sensor Resolution	13 pixels
Focus Adjustment	automatic
Special Effects	HDR
Camera Light Source	flash

IV. Our Approach

Our approach is divided into two parts:

- Image preprocessing which includes filtering, segmentation and feature selection
- Image analysis consisting of a classification algorithm based on random forests

The whole process flow of our algorithm is shown in **Figure 1** Preprocessing and feature selection is part of our **Algorithm 1** and **Algorithm 2** is used for image analysis.

Preprocessing helps in removing noise in the image using filter like mean, median. Segmentation helps in reducing the number of pixels to be analyzed. In analysis part, we used cross validation technique to first train our classifier using training images and then we test our model on 19 testing images.

A. Challenges

First, we present some challenges that need to be overcome. Recall that experimental image data is used for identify leaf stress in our problem. One of the key challenges in detecting deficient leaves was building training model. A large dataset for training results in overfitting problem and small dataset creates the problem of underfitting. Thorough analysis was done to decide on the optimal number of training image. We tested our trained model of random forests on 19 images with 14 images of leaves from tailing soil and 5 from biosolid treated tailing soil.

Amount of data: One of the main challenges in most of the image processing problem is the amount of data. Image is very

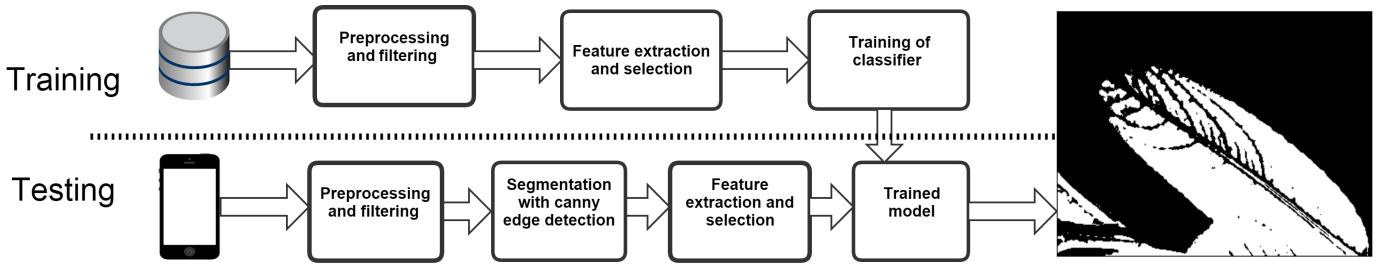

Fig. 1:

Process flow: White portion represents deficient part of leaf and black portion represents non-deficient part of leaves

large in size and to train the classifier on enormous amount of data is very hard. To train the model for each pixel on all training images takes a lot of time. To resolve this problem, we did ten fold size reduction for all the images from 4182×2322 pixels to 413×233 pixels. To further simplify the curse of large dataset for testing image, we employed segmentation to reduce the area of interest which in our case is leaves. We will discuss about the segmentation in the subsequent subsection.

Labelled data: In case of supervised learning, labelled data is required to train the classifier. There is no dataset available which contains image data with labeled data for nutrient deficiency. All the data used in our study was hand labelled under the supervision of experts from environmental sciences and converted into equivalent binary image. Binary image contains only two colors white and black which depicts two classes i.e healthy and unhealthy part of leaves. During cross validation, we checked the predicted results with the ground truth on pixel level to get accuracy matrix.

B. Noise removal using smoothing

Random variation in brightness or color information is known as noise. Accuracy of the result affects severely form the noise. Noise is not present in the actual object which is pictured. It is mainly occurs because of photo sensor which in our case is smart-phone camera. Noise sometimes occurs due to improper lighting also. Low pass filtering (also known as smoothing), is employed to remove high spatial frequency noise from a digital image.

Three filters are used in the following order for noise removal from the image.

1) ***Sharpening filter****:* Sharpening filter helps in enhancing line structures and other details in an image. In addition to original image, enhanced version of image contains scaled version of the line structures and edges in the image. Line structures and edges can be obtained by applying a difference operator which is a high pass filter on image. Combined operation is still a weighted averaging operation, but some weights can be negative, and the sum=1.

2) ***Mean filter****:* Mean filter helps in denoising the image by taking the mean of the sliding window of $n * n$ size. In our approach, we took window of size $3 * 3$ because it helps in smoothing the pixel with respect to its neighboring pixels.

3) ***Median filter****:* Median filtering is nonlinear filtering technique to remove noise. One of main characteristic of median filter is that it preserves edges while removing noise which helps during segmentation. It takes the median value instead of the average or weighted average value of pixels in the window. Median filter sort all the pixels in an increasing order, and then it takes the middle value. The output before and after applying filtering techniques is shown in **Figure 2**

Fig. 2: a) Original image b) Image after applying sharpening, mean and median filters

C. Image segmentation

The purpose of image segmentation is to divide the image in several parts with group of pixels. The goal of image segmentation is to simply and alter representation of an image. It also make image more meaningful and easy to analyze. The main purpose of using image segmentation is to find meaningful area which in our case is leaves. Our study is concerned with leaf texture, not the area outside the leaf. It help in reducing the area which we have to analyze. Regions of interest(ROI) is defined as groups of pixels having border and a particular shape such as a circle, ellipse or polygon. ROI in our study is leaf surface. In most of digital image processing pipelines, segmentation is most tricky and crucial step for image analysis. It is used mainly because of two purposes- First is to decompose the image into small part for analysis. Segmentation is very reliable if there is clear visual distinction between area of interest and the redundant area. Median filter also help to our cause by sharpening the boundary. We discarded the pixels which doesn't belongs to leaves. Secondly, segmentation presents a new representation of image. This help in organizing pixels in higher level units i.e. segments in a way that it will be more meaningful and efficient for further analysis.

Fig. 3: a) Original leaf image b. Image after applying segmentation where non-leaf pixels are removed

Identifying ROI in image depends on the uniformity and homogeneity of texture and color. Image should be clear enough to detect the similar region. Identification also depend on the different characteristics of adjacent regions of segmentation with respect to the region of interest. It is very difficult to achieve all desired properties of segmentation. Uniform and homogeneous region is full of holes and broken boundaries but our segmentation techniques take cares of these problems.

In our study, we used Canny edge detection and Sobel techniques for segmentation because they are very efficient in detecting boundary compare to other techniques. Canny edge detector get boundary segment of an intensity image. Canny operator is one of most popular edge detector. We had used σ of 0.75 for Canny operator. σ is the standard deviation of the Gaussian filter.

1) Gaussian filter: Guassian filtering is used to blur images and remove noise. Image is two dimensional. The gaussian function for two dimension is:

$$H_{ij} = \frac{1}{2\pi\sigma^2} e^{-\left((i-k-1)^2+(j-k-1)^2\right)/2\sigma^2}$$

2) Intensity Gradient of the Image: Canny algorithm uses four filters to detect horizontal, vertical and diagonal edges in the blurred image. Roberts Sobel et al edge detection operator returns a value for the first derivative in the horizontal direction G_x and the vertical direction G_y.

$$G = \sqrt{G_x^2 + G_y^2}$$
$$\theta = atan2(G_y, G_x)$$

where G can be computed using the hypot function and atan2 is the arctangent function with two arguments.

3) Sobel operator: Uses two 33 kernels which are convolved with the original image to calculate approximations of the derivative. One for horizontal changes, and one for vertical. A is source image and G_x and G_y are two images which at each point contain the horizontal and vertical derivative approximations

$$G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} * A \quad (1)$$

and

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * A \quad (2)$$

Comparison between leaf before segmentation and after segmentation is shown in **Figure 3**

D. Feature selection

As the world grows in complexity, overwhelming us with the data it generates, machine learning techniques becomes the only hope for elucidating the patterns that underlie it [21]. Manual processing of data becomes tedious as size of data grows and number of dimensions increases, so the process of data analysis need to be automated using machine learning. Feature selections plays an important role in dealing with large datasets like image data. When dealing with large digital images dataset, the computational time is increased. High collinearity and presence of noise in color bands can degrade the quality of the model. Feature selection and extraction are central issues in these situations because of the curse of dimensionality [22]. The objective of feature selection is three fold. Providing faster and more cost effective prediction, improving the prediction performance of the predictors and providing better understanding of the underlying process that generate the data. In our proposed model, we are detecting deficiency in leaves using visual characteristics. Therefore selecting right color channel or combination of color channels is very important from accuracy point of view. Some of the available color channels are RGB, HSV, YCbCr etc. The range of value for each channel pixel in case of RGB and YCbCr is 256. Range for each attribute of Hue, Saturation and value in HSV is between 0 and 1. One of the limitation of using HSV is the range of value which is very limited and decrease the accuracy of classification technique. We used various set of feature combination and calculated accuracy as given in **Table II**.

TABLE II: Comparison of various combination of color channels

Combination	Precision	Recall	Accuracy	F1-Score
RGB	0.9539	0.8976	0.9377	0.9249
R	0.9364	0.5616	0.7963	0.7021
G	0.9408	0.5786	0.8044	0.7166
B	0	0	0.5726	0
YCbCr	0	0	0.5726	0
HSV	0	0	0.5726	0

E. Classification method

One of the most crucial step of any supervised learning based solution is the classification method. We approached detecting deficiency as binary classification problem in which we have two classes: Healthy class and deficient class. During our study, we evaluated many supervised learning techniques

like Random Forests, Support Vector classification and K-Nearest Neighbor and come with Random Forests as best classification technique based on training time and prediction time. We are predicting our outcome based on pixel features of RGB color channel. We designed our algorithm based on random forests as shown in **Algorithm 2**

Algorithm 2 Algorithm for Random Forests based classification for plant stress detection

Level 1 Training:
Input: Training image data set $\vec{B}_i$
Output: Ensemble of trees $\{T_b\}_1^B$

1) Select a bootstrap sample Z^* of size N from the training data.
2) Grow a decision tree T_b to the bootstrapped image data by recursively repeating following steps for each terminal node, until minimum node size n_{min} is achieved.
   a) Select m attributes at random from the n variables.
   b) Choose the best attribute/split-point among the m.
   c) Split the node into two daughter nodes.

Level 2 Testing:
Input: Segmented test image data set $\vec{T}_i$
Output: Pixelated testing leaves with white represents unhealthy and black represents healthy leaf pixel
Classification: If $\hat{C}_b(x)$ be the class prediction of the b^{th} decision tree. Then $\hat{C}_{rf}^B(x)$ = majority vote $\{\hat{C}_b(x)\}_1^B$

$$S_i(\vec{T}_i) = \begin{cases} 1, & \text{if } T_i \epsilon C_D \\ 0, & \text{otherwise} \end{cases} \quad \forall i \in (1, N)$$

Random Forests(RF) is an ensemble supervised machine learning technique. Random forests uses decision tree as base classifier [12]. Random forests consists of set of decision trees; $h(x, \theta_i) i = 1, 2, ...,$ where the θ_i are independent identically distributed random vectors and each decision tree contribute a unit vote for the most popular class at input x. Each single tree in RF contains N number of records in the training set which is sampled at random but with replacement, from the original data, this is bootstrap sample. If there are M input features, a $m << M$ is selected as attributes at each node. The best split on these m attribute is used to split the node. RF uses information gain $IG(T, a)$ to decide splits. T denote a set of training sample for a single tree. $((\mathbf{x}), \mathbf{y}) = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_k}, \mathbf{y})$ where $(\mathbf{x})$ consist is single set of record and y is the class label. The information gain for an attribute a is as follow:

$$IG(T, a) = H(T) - \sum_{v \in val(a)} \{ \frac{|\{\mathbf{x} \in \mathbf{T} | \mathbf{x_a} = \mathbf{v}\}|}{|T|} H(\{\mathbf{x} \in \mathbf{T} | \mathbf{x_a} = \mathbf{v}\})\}$$

Here $x_a \in vals(a)$ is the value of the a^{th} attribute of example **x** The randomization is present in two ways: (1) random selection of data for bootstrap samples as it is done in bagging (2) random selection of input features for creating individual base decision trees. Robustness of individual decision tree classifier and correlation among base decision trees decides generalization error of RF classifier. Random forests works efficiently on large dataset like image dataset [12]. Ensemble technique is often more accurate than any of the single classifier in ensemble [23][13][14]. Multiple trees are induced in the forests. In our approach, nTree is set to 100 which is the total number of decision trees. Random forests do the bagging for of decision trees but we can ensemble other classifier also like neural networks. Random forests use bagging for averaging the output of each decision tree $p(c|v) = \frac{\sum_{t=1}^{T} p_t(c|v)}{T}$ The process of combining votes from all trees and selecting class with maximum votes is referred as Forest RI in the literature [12] Optimal forest size gives smoother separation and better decision boundaries. Number of output classes doesn't affect the classification accuracy. In decision trees, shallow tree results in problem of underfitting and deep tree increase the chances of overfitting. RF uses Bayesian optimization to get optimal choices for all parameters. The Generalization error (GE^*) of Random Forest is given as, [14]

$$GE^* = P_{x,y}(mar(X, Y)) < 0$$

Where $mar(X, Y)$ is Margin function. The margin function calculates the extent to which the average number of votes at (X,Y) for the correct class exceeds the average vote for any other class [24]. Here X is the predictor vector and Y is the class value.
The margin function is given as,

$$mar(X, Y) = avg_k I(h_K(X) = Y) - max_{j \neq Y} av_k I(h_K(X) = j)$$

Here $I()$ is indicator function. Expected value of margin functions is used to represent strength of random forest as,

$$S = E_{x,y}(mar(X, Y))$$

In the classification problem, margin is directly proportional to confidence. The upper bound of generalization error of ensemble classifier is represented by a function of mean correlation between base classifier and their average strength which is (S) [24]. If ρ is average value of correlation. Then upper bound for generalization error is given by:

$$GE^* \leq \frac{\rho(1 - s^2)}{s^2}$$

Our random forests based trained learning algorithm is stored in smart-phone. Everytime we take an image from the smartphone using our application, image is feed to the algorithmic technique which in return provides the healthiness of the leaf. This will save testing time and helps in getting the output very fast.

V. Results

Cross validation was used for validating our approach to determine whether or not the pixels were attempted to classify in each leaf was healthy or not. Precision, recall, accuracy and F1 score were used as accuracy measure, which are standard metrics to assess performance of classification algorithms. In **Figure 4**, we compared original testing image, ground truth image and the predicted result from Random Forests classifier. Our predicted results achieved an aggregated precision of 91%, recall of almost 75%, accuracy of 91% and F1-Score of 82% for all the images with standard deviation of 04.39%, 12.86%,

Fig. 4:

Random Forest classifier result: a. Actual image b. Black-white ground truth c. Our algorithm prediction results

TABLE III: Accuracy Matrix of Random Forests based Algorithm

	Precision	Recall	Accuracy	F-1 Score
Mean	91.288%	74.82%	91.24%	81.52%
Standard Deviation	04.39%	12.86%	05.38%	08.11%

05.38% and 08.11% respectively on sample of 19 testing leaves, as shown in **Table III**.

Subsequently, we evaluated the soil treatment on plant stress using our approach, as a notion of overall stress of the plant. After training our random forests based algorithm, we ran the algorithm on plants which were grown on tailing soil which was nutrition deficient. The same model was also applied on biosolid treated soil which contains more minerals and supported healthier plant growth. We ran our model on 5 images from biosolids amended tailings and 14 images from tailing soil plant leaves. We expressed the healthiness of plant in following terms:

$$H(L_i) = \frac{\sum_{j=1}^{M} I(L_i^j \in C_{ND}) * 100}{\sum_{j=1}^{M} I(L_i^j \in C_{ND}) + \sum_{j=1}^{M} I(L_i^j \in C_D)}$$

Where C_{ND} is class of healthy pixels and C_D is class of deficient pixels. M is the total number of pixels in leaf L_i. L_i^j denotes j^{th} pixel of leaf i.

Our feature set consists of RGB tuple. We discarded those pixels which does not belongs to leaves using segmentation technique during testing. Each pixel of the leaf is processed to check whether it belongs to deficient class or healthy class. Therefore the above equation gives us healthiness of the leaf. We evaluated the healthiness of leaves for both tailing and biosolid treated soil. Our results shown that there was a significant improvement in plant stress when treated with biosolids. These results are consistent with biosolids treated plants, as indicated by domain experts. Results are shown in **Figure 5**.

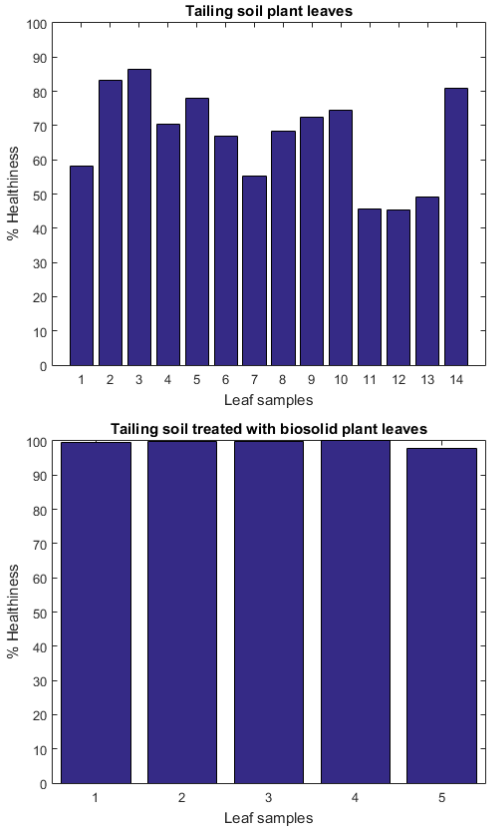

Fig. 5:

Comparison of plant leaf healthiness for tailing and biosolid treated soil

VI. Discussion

We evaluated our system on 34 leaves. Out of which 15 leaves were used for training and 19 for testing. As part of future work, we are planning to evaluate our system on large data set with different lighting settings like cloudy, foggy and daylight conditions. In our current setting, we used daylight condition of greenhouse. The beauty of our algorithms embedded in our classification techniques. Random forests take very less time in processing and evaluating the data set. Currently we evaluated our system on machine with following configuration *Intel Core i7 CPU@2.6 GHz with 16 GB RAM*. Training time of our random forest based model with nTree=100 is 46.38 seconds. For each leaf, our RF based model takes 1.1 seconds for prediction with 83.2% accuracy.

Our Smart-phone app with trained model of random forest can evaluate the image taken from cameras with every less energy and computation.

VII. Conclusion

In this paper, we demonstrate the feasibility of smart-phone cameras being used to assess stress of plants. Our methods included image filtering to remove noise, and segmentation to further improve scalability and accuracy. Our algorithm based on random forests achieved good performance in classifying the healthy vs. unhealthy portions of the plant we studied in both soils. Our future work is to generalize the above experiments to consider more plants and soils, evaluating the applications of work to urban farms, and also design smart-phone apps to further outreach our contributions to the society.

References

[1] J. Letey, "Relationship between soil physical properties and crop production," in *Advances in soil science*. Springer, 1985, pp. 277–294.

[2] J. Lipiec *et al.*, *Soil physical conditions and plant growth.* CRC Press Inc., 1990.

[3] J. Lynch and L. Audus, "Products of soil microorganisms in relation to plant growth," *CRC Critical reviews in Microbiology*, vol. 5, no. 1, pp. 67–107, 1976.

[4] A. Van Bruggen and A. Semenov, "In search of biological indicators for soil health and disease suppression," *Applied Soil Ecology*, vol. 15, no. 1, pp. 13–24, 2000.

[5] K. Tilak, N. Ranganayaki, K. Pal, R. De, A. Saxena, C. S. Nautiyal, S. Mittal, A. Tripathi, and B. Johri, "Diversity of plant growth and soil health supporting bacteria," *Current science*, vol. 89, no. 1, pp. 136–150, 2005.

[6] S. K. Prajapati and B. Tripathi, "Seasonal variation of leaf dust accumulation and pigment content in plant species exposed to urban particulates pollution," *Journal of environmental quality*, vol. 37, no. 3, pp. 865–870, 2008.

[7] A. Kessler and I. T. Baldwin, "Plant responses to insect herbivory: the emerging molecular analysis," *Annual review of plant biology*, vol. 53, no. 1, pp. 299–328, 2002.

[8] L. Chaerle and D. Van Der Straeten, "Seeing is believing: imaging techniques to monitor plant health," *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, vol. 1519, no. 3, pp. 153–166, 2001.

[9] S. Lindow and R. Webb, "Quantification of foliar plant disease symptoms by microcomputer digitized video image analysis," *Phytopathology (USA)*, 1983.

[10] T. Price, R. Gross, W. J. Ho, and C. Osborne, "A comparison of visual and digital image processing methods in quantifying the severity of coffee leaf rust (hemileia vastatrix)," *Animal Production Science*, vol. 33, no. 1, pp. 97–101, 1993.

[11] N. E. Abdullah, A. Rahim, H. Hashim, M. M. Kamal *et al.*, "Classification of rubber tree leaf diseases using multilayer perceptron neural network," in *Research and Development, 2007. SCOReD 2007. 5th Student Conference on*. IEEE, 2007, pp. 1–6.

[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] A. Krogh, J. Vedelsby *et al.*, "Neural network ensembles, cross validation, and active learning," *Advances in neural information processing systems*, vol. 7, pp. 231–238, 1995.

[14] V. Y. Kulkarni and P. K. Sinha, "Random forest classifiers: a survey and future research directions," *Int. J. Adv. Comput*, vol. 36, no. 1, pp. 1144–1153, 2013.

[15] D. Story, M. Kacira, C. Kubota, A. Akoglu, and L. An, "Lettuce calcium deficiency detection with machine vision computed plant features in controlled environments," *Computers and Electronics in Agriculture*, vol. 74, no. 2, pp. 238–243, 2010.

[16] G. A. Carter and R. L. Miller, "Early detection of plant stress by digital imaging within narrow stress-sensitive wavebands," *Remote sensing of environment*, vol. 50, no. 3, pp. 295–302, 1994.

[17] C. Bock, G. Poole, P. Parker, and T. Gottwald, "Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging," *Critical Reviews in Plant Sciences*, vol. 29, no. 2, pp. 59–107, 2010.

[18] D. Sena Jr, F. Pinto, D. Queiroz, and P. Viana, "Fall armyworm damaged maize plant identification using digital images," *Biosystems Engineering*, vol. 85, no. 4, pp. 449–454, 2003.

[19] R. Lucas and J. Davis, "Relationships between ph values of organic soils and availabilities of 12 plant nutrients." *Soil Science*, vol. 92, no. 3, pp. 177–182, 1961.

[20] A. Läuchli and S. R. Grattan, "9 soil ph extremes," *Plant stress physiology*, p. 194, 2012.

[21] I. H. Witten and E. Frank, *Data Mining Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[22] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines," *Geoscience and Remote Sensing Letters, IEEE*, vol. 4, no. 4, pp. 674–677, 2007.

[23] M. R. Kosorok, S. Ma *et al.*, "Marginal asymptotics for the large p, small n paradigm: with applications to microarray data," *The Annals of Statistics*, vol. 35, no. 4, pp. 1456–1486, 2007.

[24] R. J. Prenger, T. D. Lemmond, K. R. Varshney, B. Y. Chen, and W. G. Hanley, "Class specific error bounds for ensemble classifiers," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 843–852.

VIII. Appendix

In Table IV, we present a complete list of important chemical components of soils used in our study. Note that the presence of Lead (Pb) and Nickel (Ni) are actually harmful for our plants, and as such their decreased presence enables healthier plants and less plant stress. Some of the entries are Not Determined (ND) which are in process of chemical analysis. More chemical analysis is currently being done to determine a complete list of all chemicals, and their quantities present in both soils.

TABLE IV: Complete list of important chemical components of soils used in our study

Component	**Tailing Soil**	**biosolids**
pH	7.6	ND
CEC (meq/100g)	3.6	ND
Organic Matter(%)	0.1	56.3
Total Kjeldahl nitrogen(mg/kg)	ND	66500
Nitrate(mg/kg)	ND	69.1
Bray I P(mg/kg)	16.5	ND
Total P(mg/kg)	ND	14200
Ca (mg/kg)	473.5	ND
Mg (mg/kg)	136.5	ND
K (mg/kg)	24	3140
Cu (mg/kg)	0.999	522
Zn (mg/kg)	ND	735
Mo (mg/kg)	2.536	ND
Cr (mg/kg)	11.49	24.6
Cd (mg/kg)	13.67	ND
Pb (mg/kg)	3553	31.5
Ni (mg/kg)	70.67	22.4
Co (mg/L)	39.25	ND