Identifying Mosquito Species using Smart-Phone Cameras

Mona Minakshi, Pratool Bharti, and Sriram Chellappan

***Abstract*—Mosquito borne diseases have been amongst the most important healthcare concerns since time. An important component in combating the spread of infections in any geographic region of interest has been to identify the type of species that are prevalent in that region. As of today, dedicated personnel are assigned in most (if not all nations) to trap samples and identify them. Unfortunately, the process of identifying the actual species of mosquito is currently a manual process requiring highly trained personnel to visually inspect each specimen one by one under a microscope to make the identification. In this paper, we propose a system to automate this process. Specifically, we demonstrate results of an experiment we conducted where learning algorithms were designed to process images of captured mosquito samples taken via a smart-phone camera in order to identify the actual species. Using a total sample size of 60 images that included 7 species collected by the Hillsborough County Mosquito and Aquatic Weed Control Unit (in the city of Tampa) our proposed technique using Random Forests achieved an overall accuracy of 83.3% in correctly identifying the species of mosquito with good precision and recall. While our proposed technique will greatly benefit the state-of-the-art in species identification, we also believe that common citizens can also use our proposed system to improve existing mosquito control programs across the globe.**

I. Introduction

Mosquito borne diseases (e.g., Malaria, Dengue, West Nile Fever, and most recently Zika Fever) are amongst the biggest healthcare concerns across the globe today. To mitigate the spread of mosquito-borne diseases, it is vital to combat the spread of mosquitoes. Of critical importance in this mission is the identification of species prevalent in an area of interest. However, doing this is not at all easy. As of today, dedicated and trained professionals lay traps for mosquitoes, and pick them soon after to sort them out. Subsequently, to identify each specimen collected, it is placed under a microscope, and visually identified, which takes hours each day for all samples. Needless to say, this is a very time consuming process, and also imposes severe cognitive burden.

Our Contributions: In this paper, we design learning algorithms that process images from smart-phone cameras for automatic identification of mosquito species. Specifically, our contributions are

a). Generating a Database of 60 **image samples of** 7 **mosquito species:** In Fall 2016, we visited the Hillsborough County Mosquito and Aquatic Weed Control (in Tampa) to collect numerous samples of mosquitoes that were captured in traps set up by county personnel for species identification. Typically, when a trap is set, samples of dead mosquitoes are collected the very next day to prevent their decaying (that will complicate visual identification). Over time, the personnel there helped in visually identifying 60 different samples of mosquitoes evenly distributed from 7 different species presented in Table I. Once the species were identified, we took one image of each specimen using a Samsung Galaxy S5 smart-phone camera that revealed the maximum surface area of the mosquito specimen, and generated a database for 60 images. In Figure 1, we present one representative image from each of the 7 species we aim to classify in this paper.

Table I: Mosquito Species and Number of Samples

Species Name	Number of Samples
Culex nigripalpus (Cx Nigrip)	10
Anopheles quadrimaculatus (An Quadrim)	6
Mansonia titillans (Ma Titillans)	7
Psorophora columbiae (Ps Columbi)	10
Anopheles crucians (An Crucians)	10
Psorophora ferox (Ps Ferox)	7
Coquillettidia perturbans (Cq Peturbans)	10

b). Pre-processing Images and Features Extraction: First, we reduced the dimenisionality of our images from 2988×5322 pixels (around 3MB) per image to 256×256 pixels (around 16kB) per image to save processing overhead. To further remove noise, each image was pre-processed with median filters and smoothed out to preserve the edges and boundaries, the colors of which are important visual indicators of the species of mosquito. Naturally, we then extracted a number of RGB features from the images, and leverage the notion of Information Gain [1] to select a limited number of representative features that provide most useful information for subsequent species classification.

c). Designing a Classification Algorithm and Evaluation: Once the features are selected, we then designed an algorithm to classify each species based on the idea of Random Forest classification algorithm. Random Forest algorithms are well suited for our problem scope because they handles larger datasets (that are typical in image datasets) efficiently and quickly without causing over-fitting problems. This is because in Random Forest designs, multiple decision trees are con-

Mona Minakshi, Pratool Bharti and Sriram Chellappan are with the Computer Science & Engineering Department at University of South Florida, Tampa, FL 33620 USA, (email: mona6@mail.usf.edu, pratool@mail.usf.edu, and sriramc@usf.edu)

978–1–5386–3873–6/17/$31.00 ©

structed that serve to minimize variance of training samples during modeling which tends to avoid overfitting problems. Apart from that it is easy to train and does not assume any distribution in data [2] [3].

We conducted a comprehensive performance evaluation of our proposed system using 10-Fold Cross Validation technique. Our technique achieved a Precision of 84.5%, a Recall of 83.3%, with an overall accuracy of 83.3% in classification.

The rest of the paper is organized as follows. In Section II, important related work is discussed. In Section III, we present the data collection process. Section IV contains details about our modeling and algorithms design. We present our evaluation and results in Section V. Finally, we conclude the paper in Section VI.

II. Related Work

We present a brief overview on work related to designing newer applications from smart-phone images.

In [4] a system was developed for determining the effectiveness of soil treatment on plant stress using smart-phone cameras, wherein 34 images of leaves from two plants: willows (Salix Pentandra) and poplars (Populus deltoides x nigra, DN34) were imaged using a Samsung Galaxy smart-phone camera to detect leaf stress as a result of poor soil quality. The authors pre-processed the images using mean and median filters and extracted a number of RGB, HSV and YCbCr based features from the images. Subsequent classification via Random Forests algorithm yielded an accuracy of 91.24% in correctly detecting stressed leaves.

Greef et al. in [5] has developed a system called BiliCam which leverages smart-phone images to detect jaundice in newborns via detecting yellowing of the skin. The original RGB values of the image are transformed to YCbCr and Lab color space. Subsequently, the mean values of each color channels is determined, which resulted in a total of 9 features. The proposed algorithm is based on ensemble of different regression methods like K-NN, LARS, SVR, and Random Forests. Evaluation is done on 100 newborns and demonstrated an overall accuracy of 85% in detecting jaundice.

Other related work includes a) work in [6], where images from smart-phone cameras attached on car windshields are processed for predicting signal times and b) work in [7] where smart-phone images of drivers are used to detect tired or distracted drivers for road safety applications. These works are related in the sense that our paper also proposes an innovative application of smart-phones in the realm of detecting mosquito species, which has not been attempted before.

III. DATA COLLECTION

In this section, we present details on our data collection, and a brief description of species we have collected.

The Hillsborough County Mosquito and Aquatic Weed Control Unit (in Tampa, Florida) is a center that is dedicated to combat the spread of mosquitoes in the Hillsborough County area of Tampa. In general, it is known that upto 40 species of mosquitoes are prevalent in this county (not all of them are vectors for diseases though). Weekly, dedicated personnel from this unit lay traps for mosquitoes in areas where they are known to be in abundance, and the traps are then visited the very next day to pick up the dead moquitoes for identification. The early collection of samples is important because, once dead, they decay fast, making visual identification harder, if there are delays. Subsequently after collection, each sample is isolated, and carefully inspected under a microscope to visually identify the specimen, and statistics are recorded for futher action related to mosquito control.

Table II: Samsung Galaxy S5 Camera Features

Camera Details	Specification
Sensor Resolution	16 MP
Focus Adjustment	automatic
Special Effect	HDR
Camera Light Source	Daylight

In Fall 2016, we participated in multiple such efforts and were given 60 samples from a total of 7 different mosquito species (that were caught in traps one day earlier). Each sample was carefully identifed by experts there for us to get the ground truth data. Table I presents details on our data set, while Figures 1 (a) to (g) present one representative image of each of the seven species which we attempt to clssify in this paper. Note that all images were taken via a Samsung Galaxy S5 smart-phone under indoor conditions on the same morning with similar ambient light conditions, with the camera located one feet right above each sample. Features of the smart-phone camera used, are presented in Table II.

IV. Our Approach

In this section, we discuss our approach for classifying mosquito species from smart-phone images. There are four steps in our approach - image dimensionality reduction, noise removal, features extraction and algorithm design.

A. *Dimensionality Reduction*

The first step is dimensionality reduction. Image data can be really large, hence increasing computational complexity and time during classification, which is not desirable for our application. In our scenario, each original image is of dimension 2988×5322 pixels. Should we process this original image, then our system will entail processing more than 15 Million pixel vectors, which is beyond the processing capability of modern smart-phones and standard computing configurations. Even our analysis revealed that processing 262144 pixel vectors for image dimension of 512×512 was too complex for smart-phones. In our proposed approach, we reduced each image captured to 256×256 pixels (that we verified did not affect accuracy of classification) for subsequent processing. Note that this step reduced the image size from around 3MB to 16KB.

B. *Noise Removal*

Generally speaking, digital images are susceptible to different type of noise. Noise can occur in several ways during image capture, transmission, etc. Needless to say, with more noise, accuracy of classification becomes poor. While there are

a). An Crucians b). Ps Columbi c). Ps Ferox d). Cx Nigrip

e). Cq Peturbans f). An Quadrim g). Ma Titillans

Figure 1: Representative Sample for Each Species Classified. This Figure is best viewed in Color.

a). Ps Columbi b). An Crucians

c). Cq Peturbans d). Ps Ferox

Figure 2: Color contrast in legs of different species. This Figure is best viewed in Color.

a). An Crucians b). Cx Nigrip

c). An Quadrim d). Cq Peturbans

Figure 3: Color contrast in wings of different species. This Figure is best viewed in Color.

many techniques like sharpening filter, median filter and mean filter to reduce noise in images, in our work, we employ the median filter to reduce noise.

Specifically, median filter [8] is a nonlinear filtering technique, wherein the idea is to replace each pixel value in the window of $n \times n$ size pixel by the median of all pixel values in that particular window. This technique is widely used in digital image processing, and a very nice thing about this is that it preserves image edges while removing noise. Edge preservation is vital for our problem, since legs and wings (which are the part of edges in mosquito images) have distinct color patterns that are utilized to identify them. For instance, in Figure 2 we can see that the legs of Ps Columbi and Cq Peturbans have a combination of black and white color patches while the legs of An Crucians and Ps Ferox are purely black. Similarly in Figure 3, we can see that that the color of the wings of An Crucians and An Quadrim are dark black, while those of Cx Nigrip and Cq Peturbans are brown in color. There are distinct color patterns even in the scales of the wings, and their shapes also show visible demarcation, which are used for species identification. To wrap up, in our proposed technique, we have used median filtering technique with 3×3 pixels window size for removing the noise from our digital images.

C. Feature Selection

The third step in our proposed technique is feature selection. Feature selection is a very important part of classification. Basically, any machine learning algorithm defines a model which is essentially a relation $f(X, Y)$ between an input $X = \{x_1, x_2, ..., x_S\}$ and an output Y, built on provided training data sets, where x_i is selected input features and S is a real number. It is generally true that not every input feature x provides the same amount of information about the output Y, rather only a small subset of them $\{x_1, x_2, ..., x_m\}$

($m < S$), gives important and non-overlapping information about Y. In an ideal world where data, energy, time and processing power are in abundance, one can use all the input features including those irrelevant ones, to model the underlying function between the input and output. But in practice, there are several problems caused by the irrelevant features involved in the learning process including increased complexity, and most importantly, over-fitting problems where irrelevant features are erroneously considered important for classification during modeling.

Recall that in our proposed model, we are identifying different species of mosquitoes from image data. While the shapes do indicate a degree of difference across species, the color of samples plays the most critical role in identification, especially across legs and wings (presented earlier with examples in Figures 2 and 3). In our technique we leverage this insight via processing RGB color channels of each species, since that is most straightforward compared to techniques like HSV (hue, saturation, and value), that are better suited for situations where brightness and contrasts in images are leveraged for classification.

Since there are 256×256 pixels in our input image, the total number of initial features that we extracted was 196608 (i.e., $256 \times 256 \times 3$ for R, G and B values). This is a very large number of features. To make the problem more tractable and avoid overfitting, we employed an Information Gain [1] approach for extracted relevant features only.

The information gain for a feature F_i calculated is as follows:

$$IG\left(Tr, F_i\right) = H\left(Tr\right) - \sum_{t\ \epsilon\ F_i} p(t)H(t), \text{ where} \tag{1}$$

$$H(Tr) = -\sum_{x\ \epsilon\ m} p(x)\log_2 p(x) \tag{2}$$

Here, Tr denotes the set of training samples containing all features extracted for all image samples, and F_i denotes the i^{th} feature. The term t denotes the number of unique values for the feature F_i, and $p(t)$ is the ratio of the number of mosquitoes, for which the corresponding Feature $F_i = t$. Here, $H(Tr)$ and $H(t)$ are the entropy of the features in training set Tr and the entropy of features in the subset t respectively. The term $p(x)$ is the ratio of number of samples of a particular species x to the total number of mosquito samples in training data set Tr and m is the total number of species.

This feature selection technique above helps in achieving high accuracy and a good measure for deciding the relevance of a feature. Specifically, it gives the ranking of features based on information gain entropy value. We tried tradeoff on top 250, 500, 750, 990, 1000, 1025, 1250 and 1500 features and then selected the top 1000 features which gave us best accuracy among these. The selected features training samples served as an input vector x into Random Forest classification algorithm, a supervised learning algorithm presented next.

D. Classification Method

Random Forests (RF) is an ensemble supervised machine learning algorithm. It consists of a set of decision trees; $h(x, \theta_i)$, where $i = 1, 2, ..., n$ and θ_i are independent identically distributed random vectors. Here, x is a feature vector extracted from the smart-phone image data. Each decision tree predicts a class independently. A voting is performed on the results from each decision tree and finally the class which gets majority vote will be the final predicted class. Given a dataset that contains N training samples, each consisting of S features, the RF algorithm builds the trained model using following steps:

1) N samples are selected at random with replacement from the data set, for training the model of a particular tree.
2) K features are randomly selected from the set of available features S, where $K \ll S$.
3) Each tree will grow to its maximum size until the stopping criterion has been fulfilled.

In our implementation, the stopping criterion was a tree depth of 6, and the number of decision trees were 121 which gave us best results. Once the forest has been ensembled, testing data specimen is labeled with one of the classes $(species_1, species_2....species_7)$ by taking the majority vote: i.e., it is labeled with the class which has been selected by maximum number of trees. In the RF modeling, given a feature vector x to be classified, the conditional probabilities for each class are computed by taking the average of the conditional probabilities given by the trees constructing the ensemble. These conditional probabilities are computed as follows. Given a decision tree T, and an input feature vector x to be classified, let us denote by $v(x)$ the leaf node where x falls when it is classified by T. The probability $P(m|x, T)$ that vector x belongs to the class m, where $m \in \{species_1, species_2, ..., species7\}$ (for 7 species of interest to this paper), is determined by the following equation:

$$P\left(m|x, T\right) = \frac{n_m}{n}, \tag{3}$$

where n_m is the number of mosquito training samples falling into $v(x)$ after learning and n is the total number of mosquito training samples assigned to $v(x)$ by the training procedure. Given a forest consisting of L trees and an unknown feature vector x to be classified, the probability estimate $P(m|x)$ that vector x belongs to the species m is computed as follows:

$$P\left(m|x\right) = \frac{1}{L}\sum_{i=1}^{L} P\left(m|x, T_i\right), \tag{4}$$

where $P(m|x, T_i)$ is the conditional probability provided by the i^{th} tree and is computed according to Eq. 3. As a consequence, for vector x to be classified, the RF algorithm gives as output the vector:

$$\mathbf{p} = \left\{P\left(species_1|x\right), P\left(species_2|x\right) \ldots\ldots P\left(species_7|x\right)\right\} \tag{5}$$

The class (species) with the highest probability in the set is chosen as classified class for the i^{th} tree. The final class in our RF algorithm is the one which gets the majority vote among all activities from all decision trees in the forest [9]. The work flow of the RF algorithm with pre-processing, training and testing phase is formally shown in Algorithm 1.

Algorithm 1: RF-based Algorithm for Mosquito-Species detection

Training Image dataset = T_{tr} Testing Image dataset= T_{te}
Features extracted from Training Image dataset = F_{tr}
Features extracted from Testing Image dataset = F_{te}
Classified Species from Images= m
Probability that feature vector F belongs to Species m = $P(m|F)$

Step 1 Pre-Processing:
1) Median filters are applied to remove noise from T_{tr} and T_{te}
2) Features F_{tr} and F_{te} are extracted from processed data T_{tr} and T_{te} obtained from Step 1.

Step 2 Training:
Input: Training data set F_{tr}
Output: Random Forest model to classify different species of mosquitoes
1) Select a bootstrap sample of size N from the training data T_{tr}.
2) Grow a decision tree T using following steps
   a) Select K features at random from the set of S features
   b) Choose the best feature/ split-point based on Information Gain from Eq. 1 and 2 among the K
   c) Split the node into two daughter nodes
   d) Grow the tree to its maximum size which we fixed 6 in our case

Step 3 Prediction:
Input: Testing data set T_{te} and Trained RF model from Step 2
Output: Final Mosquito Species prediction m_{fs}
1) Select the same features F used for training the model from testing feature set F_{te}
2) Predict the species m_{fs} from the model using following equations:

for $each\ T\ in\ Forest$ **do**
$$P(m|F) = \frac{1}{L}\sum_{i=1}^{L} P(m|F, T_i)$$
end for
$$m_{fs} = \underset{i\in\{1,2.....7\}}{\operatorname{argmax}} \Big(P(m_i|F)\Big),$$
where m_i belongs to $species_i$

V. Results

Overview of Evaluation Methods: In this section, we present evaluation results of our system using 10-fold cross validation technique, which is standard for our problem scope. Very briefly, in *10-fold cross-validation* technique, the idea is to divide the entire dataset into 10 subsets, and evaluate them 10 times. Each time, one of the 10 subsets is used as the test set and the other 9 subsets are put together to form a training set. Then, the average error across all 10 trials is computed for final result.

Metrics: The results of our system are evaluated using the following metrics: Precision, Recall, F1-measure and Confusion Matrix. Based on classification of True Positives (TP), False Positives (FP) and False Negatives (FN), we have

$$Precision = \frac{TP}{TP + FP}, \tag{6}$$

$$Recall = \frac{TP}{TP + FN}. \tag{7}$$

The F_1*-measure* is the weighted average of precision and recall, and is given by,

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{8}$$

The *Confusion Matrix* (CM) is a table that shows the confusion inherent in the classification model. Each column of the matrix represents the instances in the predicted class, while each row represents the instance in the actual class (or vice-versa).

Results and Interpretations: Figure 4 presents our results. As we can see, the performance of our system in detecting the species of mosquitoes is very good. The average Recall was 83.3% and the average Precision was 84.5%. Out of the seven species we attempted to detect, the overall accuracy (F1-measure) for five of them was above 83%. Confusion Matrix of the same is shown in Figure 5.

While we believe that our results are encouraging, we do see that there is still some confusion in the ability of our technique for classification. This is because some species have very similar features to others. For instance, the abdomen of Ps Ferox and Cq Peturbans have similar looking yellow and black patches, and their legs also have yellow spots that are similar. Also, species like An Quadrim and An Crucians have similar looking color patterns in the wings that create a degree of confusion. It maybe possible that with incorporating more features, and using more samples for model training, and coupled with experimenting with emerging deep learning approaches for modeling, such confusions can be minimized, and are part of our future work.

As shown in Table III, we also implemented different machine learning algorithms for classification, and found that Random Forest technique performs the best, for the same reasons mentioned earlier in Section I.

Complexity of Execution: Our implementation of the Random Forest Classification algorithm took less than a second to predict the species of each specimen. For model development, it took a little less than 5 secs. In our results, the modeling and classification was attempted on a machine with Intel Core $i7$ CPU @2.6 GHz with 16 GB RAM configuration. Designing a smart-phone app to locally process images is part of on-going work.

VI. Conclusion

In this paper, we address an important societal scale problem, which is to detect species of mosquito samples using smart-phone cameras and image processing algorithms. We are confident that our work has important applications to mosquito control programs across the globe. Our future work lies in enhancing our database to detect more samples, work with emerging deep learning approaches to achieve superior classification accuracy, and design and deploy a real and friendly

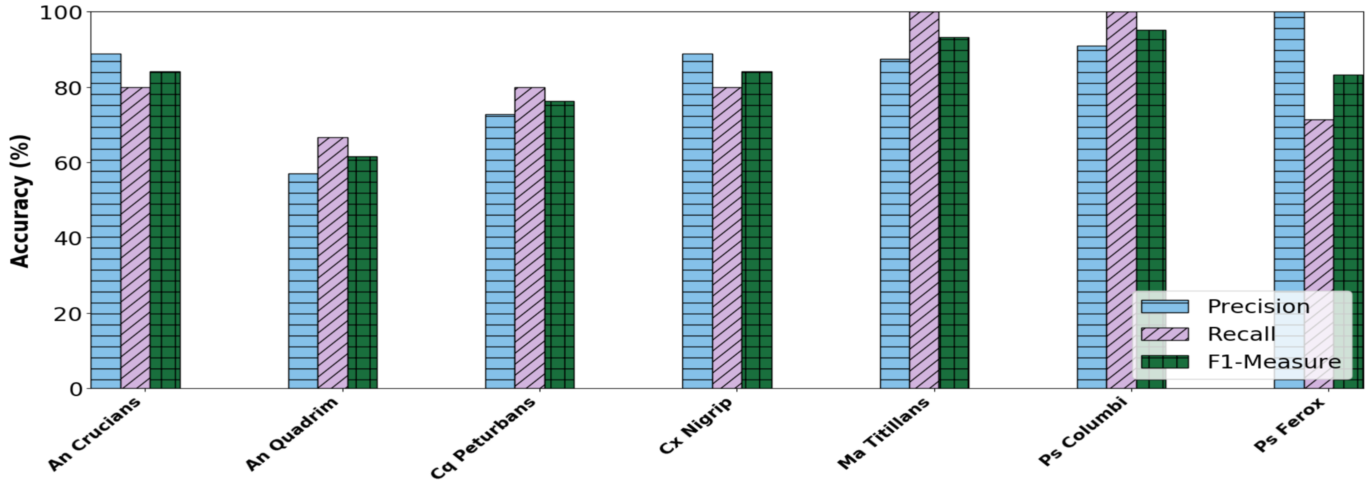

Figure 4: Precision, Recall and F_1-Measure evaluation of 10-fold cross-validation method

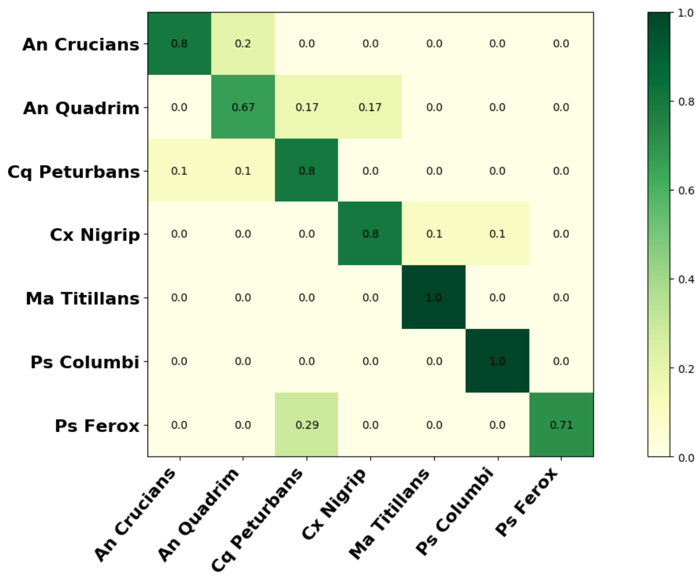

Figure 5: Confusion Matrix

Table III: Comparing Performance of Different Machine learning Algorithms

Algorithm	Accuracy
Random Forest	83.3%
KNN	63.3%
Decision Tree	56.7%
Logistic Regression	53.3%

smart-phone app that both experts and common citizens can use for image capture, identification and data sharing. We are currently also looking into incremental learning techniques for our problem in this paper. With this technique, we do not need to retrain the complete model from scratch should more images become available. Of particular interest is nearest class mean forest algorithm[10], where existing decision trees are adapted based on new images for superior and efficient classification.

ACKNOWLEDGEMENT

We thank to Hillsborough County board of Mosquito and Aquatic Weed Control, Tampa for assistance in mosquitoes sample collection. This work was partially supported by National Science Foundation under grant CNS-1205695.

REFERENCES

[1] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Information processing & management*, vol. 42, no. 1, pp. 155–165, 2006.

[2] L. Breiman, "Random forests," *Machine learning*, 2001.

[3] A. A. Montillo, "Random forests," *Lecture in Statistical Foundations of Data Analysis*, 2009.

[4] A. Panwar, M. Al-Lami, P. Bharti, S. Chellappan, and J. Burken, "Determining the effectiveness of soil treatment on plant stress using smart-phone cameras," in *IEEE MoWNeT, 2016*, April.

[5] L. de Greef, M. Goel, M. J. Seo, E. C. Larson, J. W. Stout, J. A. Taylor, and S. N. Patel, "Bilicam: using mobile phones to monitor newborn jaundice," in *ACM UbiComp, 2014*. ACM, 2014.

[6] E. Koukoumidis, L.-S. Peh, and M. R. Martonosi, "Signalguru: leveraging mobile phones for collaborative traffic signal schedule advisory," in *ACM MobiSys, 2011*. ACM.

[7] C.-W. You, M. Montes-de Oca, T. J. Bao, N. D. Lane, H. Lu, G. Cardone, L. Torresani, and A. T. Campbell, "Carsafe: a driver safety app that detects dangerous driving behavior using dual-cameras on smartphones," in *ACM UbiComp, 2012*. ACM.

[8] Wikipedia, "Median filter - wikipedia, the free encyclopedia," 2016, [Online; accessed 23-May-2016]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Median_filter&oldid=721681480

[9] L. Rokach and O. Maimon, "Decision trees," in *Data Mining and Knowledge Discovery Handbook*. Springer US, 2005, pp. 165–192.

[10] M. Ristin, M. Guillaumin, J. Gall, and L. Van Gool, "Incremental learning of ncm forests for large-scale image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3654–3661.