An Empirical Study on Machine Learning Models for Potato Leaf Disease Classification using RGB Images

Soma Ghosh, Renu Rameshan and Dileep A. D.
Indian Institute of Technology, Mandi, Himachal Pradesh, India

Keywords: Potato Leaf Disease Classification, Early Blight, Late Blight, Healthy Leaves, Salient Region based Segmentation, Gabor Filter Bank, Focus/ Defocus Cue, Convolutional Neural Network (CNN), Fully Connected Neural Network (FCNN), Support Vector Machine (SVM).

Abstract: In this work, an empirical study is conducted on classification models built using RGB images of potato leaves. A series of experiments are done by training convolutional neural network (CNN) and support vector machine (SVM) using images captured in laboratory and field conditions and processed samples of images captured in field. A salient region based segmentation algorithm is devised to generate processed version of the images captured in field which performed well with respect to manually segmented ground truth of the dataset. Severe inconsistencies are observed in experimental results, particularly when train and test samples of models are similar images but captured under different environmental conditions. Following the analysis of obtained results, we come up with a set of clear directions to create an image dataset, which can lead to a reliable classification accuracy.

1 INTRODUCTION

In spite of commendable agricultural advancement, loss of yield due to pests and fungal infections in plants causes havoc in agriculture and eventually the economy. Early detection and diagnosis of diseases can prevent an epidemic, but this process is still heavily dependent on manual expertise and laboratory based diagnosis which is time consuming, laborious (at least 1–2 days for sample harvest, processing and analysis (Martinelli et al., 2015)) and sometimes erroneous for non-native diseases (Ngugi et al., 2020).

An automatic disease detection software implemented in unmanned aerial vehicles (UAV) or handheld mobile devices can play a significant role in preliminary diagnosis of diseases in the fields, particularly for wide area plantation where manual surveillance is next to impossible and/or for remote areas with limited logistics. Usually disease spots on plant leaves are clearly visible in an RGB image and since mobile cameras are good enough to capture such images, using RGB images of plant leaves to train a standard machine learning models is a cost-effective solution to create such automated systems.

However an RGB image (referred to as "image" now onward) based model exhibits robustness and guaranteed performance only when it learns precise features related to the presence and absence of disease spot/s on the leaves properly. Here precise means, the features should be learnt from that region of the image where disease spot is present and proper means, the features should be discriminative. Extraction of relevant and discriminative features from images of plant leaves is still a very challenging task and an open research topic till date.

In this work an empirical study is conducted on machine learning models trained using RGB images of potato leaves to classify these images into three classes - *early blight*, *late blight* and *healthy leaves*. Assuming that issues related to image quality like background in images, uneven illumination, defocus, low inter-class and high intra-class variation of sample images etc. and environment of sample collection may impact the behaviour of image classifiers, we have experimented mainly with datasets than the architecture of classification models.

Different classifiers or models are built by varying the following types of dataset: 1) original *in-field* dataset, 2) original *lab-prepared* dataset, 3) augmented *in-field* dataset, 4) segmented *in-field* dataset, 5) patches extracted from *in-field* samples and 6) vector representation of *in-field* images based on the confidence scores generated by a patch based classifier. Though primary architecture of the models is convo-

Ghosh, S., Rameshan, R. and D., D.
An Empirical Study on Machine Learning Models for Potato Leaf Disease Classification using RGB Images.
DOI: 10.5220/0010234805150522
In *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2021)*, pages 515-522
ISBN: 978-989-758-486-2

lutional neural network (CNN), a support vector machine (SVM) is also trained using vector representation of *in-field* images.

Average accuracies of the models range from 98.60% (achieved by model trained on *lab-prepared* samples and evaluated on corresponding test samples) to 40.30% (achieved by model built using *lab-prepared* test samples and evaluated by segmented *in-field* samples). Inconsistent experimental results reveal the adverse effect of issues related to image quality mentioned earlier, on the learning of desired features capturing the presence or absence of disease spot/s on leaves from images. In summary, our contributions in this work are:

1. A segmentation algorithm is devised based on focus/defocus cue of the images of *in-field* dataset which achieved 77.32% overlap index, 81.3% and 94.22% precision and recall, respectively with respect to the manually created ground truth of the same dataset.
2. Patches extracted from original *in-field* images are automatically class-labelled using this segmentation algorithm.
3. Design of patch based and vector representation based classifiers. Results of these show the adverse effect of a) low inter-class and high intra-class variation of samples and b) improperly captured images on the classification performance.
4. Abysmal testing and cross-testing results of the models prove that, though learnt features seem to be discriminative enough, networks are not learning precise features relevant to disease spot/s and healthy leaf parts as desired.

The organisation of the paper is as follows: section 2 contains the related works in automatic plant disease detection using machine learning techniques. Section 3 and 4 contain the experimental method and observation and analysis, respectively. Section 5 presents the conclusions of the work.

2 RELATED WORK

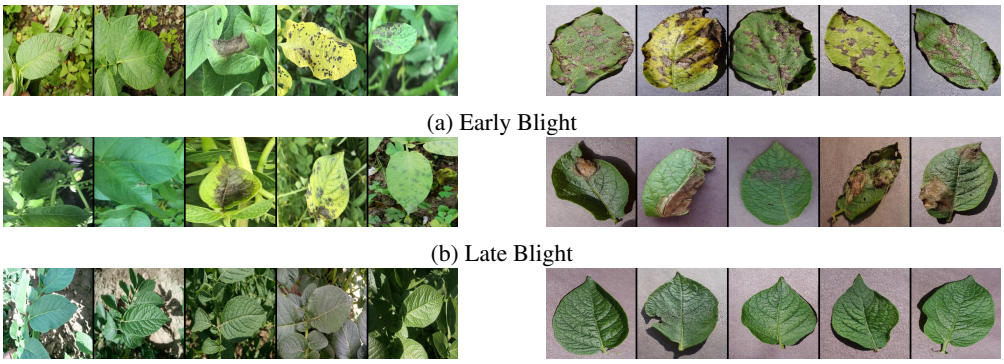
We are reviewing only those works from literature which deals with RGB image based automatic plant disease detection systems. Several methods are reported using various state-of-the-art machine learning models trained on RGB images of different kinds of crop and vegetable plant leaf images (Ngugi et al., 2020), (Saleem et al., 2019) and (Kaur et al., 2018). All these works mainly compared performance of conventional machine learning models like SVM and /or well-known CNN models like AlexNet (Krizhevsky et al., 2012), VGG16 (Simonyan and Zisserman, 2015), InceptionV3 (Szegedy et al., 2015), FRCNN (Ren et al., 2015), SSD (Liu et al., 2016) etc. Most of these methods used a *lab-prepared* dataset - *Plant Village* (Mohanty et al., 2016) and reported good classification performance (average accuracy is 96%). Some works are also reported using a few hundred to few thousand of *in-field* samples with accuracies ranging from 89% to 98%.

However in these works, adaptability and effectiveness of the classifiers are not unquestionably established for images of similar plant species but captured in different real-world situations and these works lack in justification of selection of the used CNN architectures. Also the analysis and reasoning of varying performance by compared models on same dataset are not reported thoroughly. Methods using *in-field* data also lack in analysis and justification of achieving high accuracy with such less real-world data. Few methods used well-known CNN visualisation methods and reported the highlighted disease spots in *lab-prepared* images (Ngugi et al., 2020). But again quantitative and qualitative measures of highlighted image regions and detailed analysis of their impact on model's performance is not reported explicitly. Hence these works are not being reviewed in detail.

As stated earlier, the key factor in creating an efficient plant leaf image based disease detection system is proper extraction of discriminative and relevant features of disease spots and leaves from images. Hence, emphasis of these experiments should be to analyse the suitability of sample images for learning desired features by observing the impact of quality of images on performance of the models. Also, for proper validation of plant disease classification models, experiments should be conducted using both *lab-prepared* and *in-field* images on models of similar architecture.

3 EXPERIMENTAL METHOD

In this work, the emphasis is on studying the impact of various data related issues on the behaviour of machine learning models trained to classify images of potato leaves with or without diseases. So, we started the experiments by fine-tuning a pre-trained CNN using images captured in widely varying conditions, namely *in-field* and *lab-prepared* datasets as described in section 3.1.1. The pre-trained VGG16 model (Simonyan and Zisserman, 2015) is chosen for fine-tuning due to its relatively smaller size and excellent performance on ImageNet dataset (Deng et al.,



(a) Early Blight

(b) Late Blight

(c) Healthy Leaves

Figure 1: Sample images from two datasets - first column corresponds to the CPRI dataset (*in-field*) and second column to the PV dataset (*lab-prepared*).

2009). For both types of dataset, different fine-tuned models are built by changing the numbers of higher convolutional layers of VGG16 as described in section 3.2.1.

While testing the trained models we observed a large difference between the performances of models trained on *lab-prepared* images and models trained on *in-field* samples. Since the number of samples are less, the network may not learn features which faithfully represent the complex *in-field* dataset. Hence this dataset is augmented using geometrical transformations. Classifiers built using the augmented dataset, performed better than their unaugmented counterparts with an average improvement of 5% in classification accuracies; yet this performance is inferior to the performance of models trained on *lab-prepared* samples.

Assuming that this is due to the adverse effect of background in the *in-field* images, a segmentation algorithm is devised for *in-field* images and similar experiments are conducted using segmented *in-field* dataset. Observing the poorer performance by these models, we designed patch based vector representation of *in-field* images to force the model to learn discriminative features of disease spots, healthy parts of leaves and background and to take classification decision according the presence or absence of disease spots. For uniformity across all the experiments, input size and values of hyper-parameters are kept identical as described in experimental setup (section 3.3).

Excellent performance of models trained on *lab-prepared* images and gradual degradation in results of models trained on processed *in-field* images lead us to perform cross-testing of the models trained on whole images. *By cross-testing we mean, testing generalization ability of the models on test samples taken from similar but different datasets.* Observations and analysis of the experiments follow in section 4.

(a) (b) (c) (d)

Figure 2: (a) original image, (b) ground truth, (c) segmentation mask and (d) segmented image.

Table 1: Evaluation metrics of segmentation mask.

Overlap Index	Precision	Recall	F-measure
77.32%	81.3%	94.22%	86.69%

3.1 The Datasets

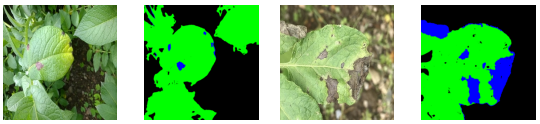
3.1.1 Original Datasets

1. ***In-field* or CPRI Dataset:** set of images provided by Central Potato Research Institute (CPRI), Shimla, India, captured in potato plantation sites (1). This dataset contains total 3387 images among which 113 images are of *Healthy Leaves* class, 1781 and 1493 images are of *Early Blight* and *Late Blight* class, respectively.
2. ***Lab-prepared* or *PlantVillage* (PV):** a public dataset (Mohanty et al., 2016) (1). This dataset contains a total of 2152 images captured in laboratory - 152 images in *healthy leaves* class and 1000 images in each of the disease classes.

Both the datasets are divided into training-validation-test sets in the ratio 8:1:1.

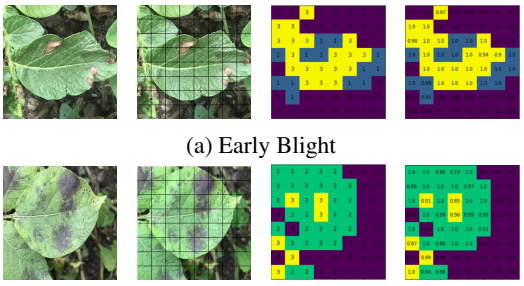
3.1.2 Derived Datasets

1. **Augmented Dataset (augD):** Training samples of CPRI dataset are geometrically transformed by random flip and random rotations (angle variation = 20°). Total 21000 images are created with 7000 images of each class.



(a) Early Blight (b) Late Blight

Figure 3: Colour-coded saliency map - first and third are original images; second and fourth images are the salient maps. Blue colour denotes disease spots on the leaves and green colour marks the healthy parts of leaves.



(a) Early Blight

(b) Late Blight

Figure 4: Patch based representation of an image: from left to right - original image, image divided into patches, patch label map and confidence score map. (Patch labels in (c)-0:*background*, 1:*early blight*, 2:*late blight* and 3:*healthy*).

2. **Segmented Dataset (segD):** CPRI images are segmented based on the concept that, focused regions have more edge information than defocused or blurred regions. A binary saliency map is created for an image by its local frequency analysis using two-dimensional (2D) Gabor filter bank. A green hue mask is also created to further leverage the extraction of focused leaves part with disease spots only. Figure 2 depicts the resulted segmentation mask and performance metrics of the devised segmentation algorithm evaluated using manually segmented ground truth images of CPRI dataset is in Table 1.
3. **Image Patch Dataset:** Each image in the two disease classes of both *in-field* and *lab-prepared* dataset is divided into overlapping patches and categorized in four classes - *background*, *early blight*, *late blight* and *healthy*. The colour coded saliency map (Figure 3) generated by the segmentation algorithm devised earlier is used to discriminate between green leaf part and the disease spots. Patches extracted from train-validation-test sets of original images are used for respective tasks only.
4. **Vectorized Image Dataset:** Each image in *in-field* dataset is divided into non-overlapping patches and fed to the patch based classifier to generate patch-wise classification scores for each image (Figure 4). As spatial position of the classified patches is no more significant for classification of the original image; each image is represented by four class-wise normalised aggregated confidence scores, corresponding to *background*, *early blight*, *late blight* and *healthy* class, respectively.

3.2 The Classifiers

3.2.1 Fine-tuned Models

To built the fine-tuned models, pre-trained VGG16 is trained along with one fully connected (FC) layer of 1024 nodes and one classification layer of three nodes to produce classification scores. For each fine-tuned model, FCs are reinitialised.

1. **FCNN:**
   FCNN is trained on features extracted from the images of training and validation sets of a dataset using pre-trained VGG16. Models built using original CPRI and PV samples are named as FCNN-CPRI and FCNN-PV, respectively. FCNN-augD, FCNN-segD are built from augmented and segmented CPRI training samples, respectively.
2. **FT13:**
   These models are created by fine-tuning last convolutional layer of pre-trained VGG16 along with an initialised FCNN. Depending on the used dataset, the built classifiers are referred to as FT13-CPRI, FT13-PV, FT13-augD and FT13-segD.
3. **FTB5:**
   Last three layers in fifth block (B5) of pre-trained VGG16 are fine-tuned with a redefined FCNN. Four such classifiers are built and referred to as FTB5-CPRI, FTB5-PV, FTB5-augD and FTB5-segD.

3.2.2 Trained from Scratch Models

The shallow networks (SNet) are sequential networks having filter of size 3×3 in each convolutional layer and one maxpool layer after each pair of convolutional layers. There is one fully connected (FC) layer with 1024 nodes and the classification layer to produce three class scores. Except the input layer, input of all the convolutional and fully connected layers of SNets are batch normalised. l_2 kernel regularisation is also applied to all the layers of SNets. Depending on the used dataset, number of layers and number of filters in each layer are varied:

Table 2: Testing results of the classifiers (Description of cell contents and naming convention of classifiers and datasets are as per section 4).

Train-Test set	Classifier			
	FCNN	FT13	FTB5	SNet
CPRI-CPRI	88.87(0.9051, 0.7440, 0.7896)	85.67(0.8606, 0.7223, 0.7600)	85.07(0.8289, 0.6645, 0.6853)	-
PV-PV	98.60(0.9712, 0.9711, 0.9711)	98.14(0.9869, 0.9489, 0.9663)	96.74(0.9872, 0.9300, 0.9548)	-
segD-segD	84.48(0.8155, 0.7848, 0.7985)	85.97(0.8155, 0.8459, 0.8287)	80.90(0.7556, 0.7556, 0.7556)	88.06(0.8205, 0.8728, 0.8433)
augD-augD	88.66(0.8508, 0.8280, 0.8387)	88.96(0.8683, 0.8308, 0.8478)	88.06(0.8123, 0.8429, 0.8263)	92.84(0.9110, 0.8789, 0.8927)

Table 3: Cross-testing results of the classifiers (Description of cell contents and naming convention of classifiers and datasets are as per section 4).

Train-Test set	Classifier			
	FCNN	FT13	FTB5	SNet
CPRI-PV	44.19(0.3218, 0.3356, 0.2467)	46.05(0.1995, 0.3300, 0.2190)	41.40(0.1697, 0.2967, 0.2007)	-
PV-CPRI	53.73(0.2905, 0.3226, 0.2881)	49.25(0.2721, 0.3026, 0.2818)	51.94(0.3670, 0.3734, 0.3556)	-
PV-segD	46.87(0.3271, 0.3462, 0.3216)	45.08(0.3158, 0.3338, 0.3094)	40.30(0.4078, 0.3625, 0.3385)	-
segD-PV	46.98(0.4575, 0.4122, 0.4226)	45.12(0.4471,0.4367,0.4321)	58.60(0.5857,0.5522,0.5635)	65.58(0.5249,0.4700,0.4372)
augD-PV	48.84(0.4323,0.5200,0.3996)	48.84(0.4786,0.5011,0.4227)	53.02(0.5058,0.5500,0.4688)	48.37(0.4347,0.4600,0.4449)

1. The SNets built using augmented and segmented CPRI samples (SNet-augD and SNet-segD) have ten convolutional layers, five max-pool layers and total 736 convolutional filters as per following layer-wise arrangement: $32-32-32-64-64-128-128-128$.
2. The SNet built on patches (SNet-patch) has six convolutional layers, three max-pool layers and total 272 convolutional filters as per following layer-wise arrangement: $16-32-32-64-64-128$.

3.2.3 Support Vector Machine

An SVM is trained using the vectorized images of CPRI dataset. Kernel function is the radial basis function (RBF) with trade-off parameter (C) $= 10$ and $\gamma = 0.1$, set empirically.

3.3 Experimental Setup

For fine-tuning the models, images from all the datasets are resized to 224×224 pixels. Input image size for SNet-augD and SNet-segD is 256×256 pixels and for SNet-patch the size is 128×128 pixels. All input images or patches are with three colour channels (RGB).

Models are trained for 200 epochs using original CPRI and PV samples and models built from derived datasets are trained for 500 epochs. In all the experiments, categorical cross-entropy loss is minimized using RMSprop optimizer with learning rate 10^{-4}. The hyper-parameter (λ) for l_2 regularization in SNet-augD and SNet-segD is set to 10^{-3} and for SNet-patch it is 10^{-4}.

4 OBSERVATION AND ANALYSIS

Severe inconsistencies are observed in the performance of the models presented in Table 2 to Table 10. *Class names are abbreviated in all the tables as: EB - early blight, LB - late blight and HL - healthy leaves. The "model-dataset" naming convention of classifiers are as per section 3.2. Each cell in the table represents accuracy(mean precision, mean recall, mean f-measure) for the corresponding model.* Lowest testing accuracy of the models trained on PV images is 96.74% (by FTB5 corresponding to PV-PV train-test pair in second row of Table 2), whereas the highest cross-testing accuracy of these models is 53.73% (by FCNN trained on PV images and tested on the original CPRI test set in second row of Table 3). This anomaly is more perceivable in the confusion matrices of these models. Class-wise samples are perfectly classified in testing as can be seen in Table 4, whereas in cross-testing most of the CPRI test images are classified as *early blight* by PV models (Table 5). On the other hand, most of the segmented CPRI images are classified as *late blight* by PV models (Table 6).

Similarly for models trained on original CPRI images, the lowest testing accuracy is 85.07% (by FTB5 corresponding to CPRI-CPRI train-test pair in first row of Table 2), but highest cross-testing accuracy is 46.05% (by FT13 trained on CPRI images and tested on PV test images in first row of Table 3). Models trained on augmented and segmented CPRI images followed the same trend. As can be seen in testing confusion matrices (Table 7 and Table 9) of these models that, test images of corresponding datasets are moderately classified. But in cross-testing, most of the PV test samples are classified as *early blight* (Ta-

Table 4: Testing confusion matrices of the classifiers trained on original CPRI and PV datasets (Class name abbreviations and model-dataset naming conventions of classifiers in present and following tables are as per section 4).

Actual	Predicted																	
	FCNN			FT13			FTB5			FCNN			FT13-PV			FTB5-PV		
	Train-Test dataset: CPRI-CPRI									Train-Test dataset: PV-PV								
	EB	LB	HL	EB	LB	HL	EB	LB	HL	EB	LB	HL	EB	LB	HL	EB	LB	HL
EB	178	18	0	175	20	1	172	24	0	99	1	0	99	1	0	99	1	0
LB	15	104	0	15	104	0	9	109	1	0	99	1	1	99	0	0	100	0
HL	7	4	9	6	6	8	12	4	4	0	1	14	0	2	13	0	3	12

Table 5: Cross-testing confusion matrices of the classifiers trained on original CPRI and PV datasets.

Actual	Predicted																	
	FCNN			FT13			FTB5			FCNN			FT13			FTB5		
	Train-Test dataset: CPRI-PV									Train-Test dataset: PV-CPRI								
	EB	LB	HL	EB	LB	HL	EB	LB	HL	EB	LB	HL	EB	LB	HL	EB	LB	HL
EB	94	6	0	98	2	0	88	12	0	165	31	0	145	51	0	154	37	5
LB	99	0	1	99	1	0	99	1	0	104	15	0	99	20	0	94	16	9
HL	9	5	1	10	5	0	14	1	0	14	6	0	8	12	0	14	2	4

Table 6: Cross-testing confusion matrices of the classifiers trained on PV datasets and tested on segmented CPRI dataset.

Actual	Predicted								
	FCNN			FT13			FTB5		
	EB	LB	HL	EB	LB	HL	EB	LB	HL
EB	85	111	0	81	115	0	52	140	4
LB	47	72	0	49	70	0	36	80	3
HL	6	14	0	6	14	0	1	16	3

ble 8 and Table 10). In fact, average cross-testing accuracy of all the models is below 50%, even cross-testing performances of well-behaved models trained on PV images are similar to the models trained on CPRI images.

So cross-testing results of these models implies that, apparent class discriminative features learnt by the networks from their respective datasets are not precisely relevant to the disease spots and healthy parts of leaves, which are common to the images of all the datasets irrespective of their background. So, if some features related to presence or absence of disease spot/s on the leaves are actually learnt by the models, classifiers might have shown bias towards a certain set of samples in cross-testing, but would not show such appalling anomalies. Particularly, classifiers trained on PV images should have detected common features from segmented images and vice versa, as images from these two datasets are quite similar in appearance (second column of Figure 1 and fourth column of Figure 2).

To investigate these inconsistencies both PV and CPRI datasets are observed visually. It is noticed that, PV images have visual difference between samples of three classes (Figure 1). Majority of *early blight* images have many small disease spots on the leaf, whereas *late blight* images have one or two big spots on a leaf. All the *healthy leaves* images contain one very clean and well shaped leaf compared to the images of the other two classes.

In contrast to the PV images, CPRI images of all the classes have cluttered background comprised of weeds, ground patches with wide illumination variation. Foreground of these images contains some prominent leaves having prevalent uniformity of shape, texture and colour across all the classes with or without noticeable disease spot/s. Moreover, overall appearance of CPRI images are varying image-wise rather than class-wise.

From these observations it can be inferred that, learnt features are capturing overall organization of foreground and background in the images. This hypothesis is well supported by the wide difference of testing accuracies of models trained on PV images and original and augmented CPRI images. As class-wise combination of foreground and background are quite distinguishable in PV images than CPRI images, class discriminative patterns learnt from PV images are better separable by classifiers than those of CPRI images.

The degraded performance of segmented CPRI image based classifiers also provide strong indication to the importance of background-foreground combination as class discriminative features of the datasets; as without background, segmented images appeared to be further less discriminative, class-wise. As overall combination of foreground and background in images is bound to vary with datasets, all types of models performed equally poorly in cross-testing.

In continuation to the observation of dataset, it must be noticed in Figure 1 that, there is no visi-

Table 7: Testing confusion matrices of the fine-tuned models trained on segmented and augmented CPRI dataset.

Actual	Predicted																	
	FCNN			FT13			FTB5			FCNN			FT13			FTB5		
	Train-Test dataset: augD-augD									Train-Test dataset: segD-segD								
	EB	LB	HL	EB	LB	HL	EB	LB	HL	EB	LB	HL	EB	LB	HL	EB	LB	HL
EB	180	14	2	180	13	3	174	15	7	171	22	3	166	25	5	167	25	4
LB	14	103	2	15	104	0	12	106	1	19	99	1	12	106	1	25	91	3
HL	4	2	14	5	1	14	4	1	15	5	2	13	2	2	16	4	3	13

Table 8: Cross-testing confusion matrices of the fine-tuned models trained on segmented and augmented CPRI dataset.

Actual	Predicted																	
	FCNN			FT13			FTB5			FCNN			FT13			FTB5		
	Train-Test dataset: augD-PV									Train-Test dataset: segD-PV								
	EB	LB	HL	EB	LB	HL	EB	LB	HL	EB	LB	HL	EB	LB	HL	EB	LB	HL
EB	92	6	2	94	6	0	88	11	1	63	36	1	64	35	1	70	29	1
LB	84	4	12	92	3	5	70	17	13	62	34	4	67	27	6	47	49	4
HL	6	0	9	7	0	8	5	1	9	3	8	4	5	4	6	3	5	7

Table 9: Testing confusion matrices of the SNets trained on segmented and augmented CPRI datasets.

Actual	Predicted			Predicted		
	SNet-augD			SNet-segD		
	EB	LB	HL	EB	LB	HL
EB	182	13	1	172	17	7
LB	4	114	1	12	106	1
HL	5	0	15	3	0	17

Table 10: Cross-testing confusion matrices of the SNets trained on segmented and augmented CPRI datasets and tested on PV dataset.

Actual	Predicted			Predicted		
	SNet-augD			SNet-segD		
	EB	LB	HL	EB	LB	HL
EB	53	47	0	41	59	0
LB	41	45	14	0	100	0
HL	5	4	6	0	15	0

ble difference between disease spots of two classes in CPRI dataset. Whereas, shape and size of the disease spots on the leaves of a single class is varying widely, ranging from barely notable smudge to a patch covering almost the entire leaf. Such wide intra-class and narrow inter-class variation of samples hinder the learning of class discriminative features. Performance of patch based classifier is evidence to this fact.

Classification accuracy of this network is quite good (93.78% in first column of Table 11), however it is confused between two disease classes (Table 12). Approximately 96% of misclassified patches from *early blight* class belong to *late blight* class and around 82% of misclassified patches from *late blight* class is marked as *early blight* patches. Whereas, less than 10% of misclassified patches from both the disease classes are in *background* or *healthy* patch class.

The patch based vector representation of whole images in CPRI dataset shows other confusions due to unevenly illuminated leaf parts and tiny and blurry disease spots on leaves (Figure 5). As majority of images in all the classes have such drawbacks, vector representations of the whole images are erroneous causing the lowest classification score by the model trained on vector represented images (in second column of Table 11). Moreover, as per potato disease literature ((Weingartner, 1981), (Thurston and Schultz, 1981)), two types of disease spots have distinctive patterns, which are not perceivable in most of the images in any of the two datasets.

Table 11: Testing results of the classifiers trained on patches and vector representation of CPRI images.

SNet-patch	SVM
93.78(0.9391, 0.9420, 0.9401)	80.90(0.7000, 0.8000, 0.7100)

Table 12: Confusion matrices of the classifiers trained on patches and vector representation of CPRI dataset. (EB: early blight, LB: late blight, HL: healthy leaves and BG: background.)

Actual	Predicted						
	SNet-patch				SVM		
	EB	LB	HL	BG	EB	LB	HL
EB	3606	424	9	11	152	11	33
LB	187	2833	23	17	6	104	9
HL	12	22	2913	53	4	1	15
BG	36	8	14	2942			

5 CONCLUSIONS

In this work, we explored several ways to build an effective classifier for potato leaf disease using RGB images. Based on our observations, we recommend to create a large dataset having following properties to create an effective and efficient image based automated system - 1) collected samples should have minimum interference from background, 2) a single image is desired to have even illumination throughout, 3) disease spots on the leaves must be entirely in focus so that distinctive patterns of two types of disease

(a) Early Blight (b) Late Blight (c) Healthy Leaves

Figure 5: Misinterpretation of patches due to defocus and/or uneven illumination and/or unnoticeable disease spots. First, third and fifth are original images ; second, fourth and sixth images are corresponding patch label maps. (Patch labels - 0: background, 1: early blight, 2: late blight and 3: healthy.)

spots are intelligible in image, 4) in case of multiple spots of varying sizes on a cluster of leaves, multiple images must be captured and 5) samples of a particular disease must be labelled as per their maturity phase, as appearance of disease spots differs widely with time. These properties are expected to provide clarity to region of interest and to ensure sufficient class separability of the sample images. A robust segmentation algorithm is also required to be devised to extract a single leaf with disease spots or a healthy leaf only from the images captured in field.

Our observations suggest that, classification accuracy alone is not a good performance metric for this type of systems. We deduce that, quantitative and qualitative measure of features learnt by the models can only establish the trustworthiness of such models with some guarantee. We are carrying out an exhaustive analysis of features learnt by the models to prove our hypothesis that features represent the overall organisation of an image rather than leaf and disease region alone.

ACKNOWLEDGEMENTS

This research work is a part of the project *FarmerZone*, sponsored by Department of Biotechnology, Govt. of India. We thank Central Potato Research Institute (CPRI), Shimla, India for providing the dataset from field.

REFERENCES

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Kaur, S., Pandey, S., and Goel, S. (2018). Plants disease identification and classification through leaf images: A survey. *Archives of Computational Methods in Engineering*, 26:507–530.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. (2016). Ssd: Single shot multibox detector. volume 9905, pages 21–37.

Martinelli, F., Scalenghe, R., Davino, S., Panno, S., Scuderi, G., Ruisi, P., Villa, P., Stroppiana, D., Boschetti, M., Goulart, L. R., Davis, C. E., and Dandekar, A. M. (2015). Advanced methods of plant disease detection. a review. *Agron. Sustain. Dev.*, 35:1–25.

Mohanty, S. P., Hughes, D. P., and Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Front. Plant Sci.*, 7:1419.

Ngugi, L. C., Abelwahab, M., and Abo-Zahhad, M. (2020). Recent advances in image processing techniques for automated leaf pest and disease recognition – a review. *Information Processing in Agriculture*.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc.

Saleem, M. H., Potgieter, J., and Arif, K. M. (2019). Plant disease detection and classification by deep learning. *Plants*, 8.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Thurston, H. and Schultz, O. (1981). Late blight. *Compendium of Potato Diseases*, pages 41–42.

Weingartner, D. (1981). Early blight. *Compendium of Potato Diseases*, pages 43–44.