Project Report

Classification of New Mexico Chile pepper plant disease using Multilayer Neural Network model.

Introduction

The severity of diseases caused by pathogens varies from mild symptoms to decline of the infected plants, depending on the aggressiveness of the pathogen, host resistance, environmental conditions, duration of infection and other factors. Plant disease symptoms vary with the infecting pathogen and the infected part and can include leaf spots, leaf blights, root rots, fruit rots, fruit spots, wilt, dieback and decline.

Worldwide, per capita availability of food is projected to increase around 7 percent between 1993 and 2020, from about 2,700 calories per person per day in 1993 to about 2,900 calories. This is gradually becoming a mere dream because of plant disease which reduces yields. This implies plant disease have both direct and indirect impact on health, food security and economic growth of every nation. Since plant diseases are strongly influenced by environmental factors, it will be unrealistic to talk about all plants and all diseases (they are heterogenous). Chile is one of the most popular and promising grown plant in New Mexico.

Motivation: New Mexico is the nation's largest Chile pepper grower, followed by California, Arizona and Texas. It is obvious that Chile farm and produces are insufficient as about 80 percent of the Chile peppers consumed in the United States are imported, largely due to lower hand labour costs, lack of adequate funds and disease control.

Accurate and early identification is essential in tracking plant disease. Initially, the identification of plant diseases solely relies on visual examination. The process is not efficient and is also prone to human error. For a trained computer with classifier algorithms, diagnosing plant disease becomes easy and efficient. Machine learning algorithms recognize plant disease type, severity, and so on by sorting through hundreds to thousands of photos of diseased plants (Samuel, 2017).

The goal of this project is to build a robust model and application using machine and deep learning approach for image classification of Chile pepper plant disease. This will be achieved using keras package with python (by Francois Chollet and J.J. Allaire, 2018) with TensorFlow (by google, 2015 and updated January 2018) as backend. The main objectives are to use three techniques to address overfitting and train a robust model.

Overfitting: Overfitting is one of the big problems of both machine and deep learning models especially in the cases of limited training data samples. It occurs when model performance is very high on training dataset and very low on either validation, test or both. This study tackles the problem by implementing the technique of image augmentation and dropout regularization. Biasness was addressed by using three-way random data-splitting method (Training, Validation, and Testing), 13% of the whole data set removed entirely and stored in a different folder for final test purposes. The remaining were later split into 70-30% training–testing part.

Methods

Data set Description: The considered pathogenic disease affect Chile plant stem and thereby block both water and nutrient intake for leaves. This study recognizes area base effect on intensity values of images, thereby uses images peculiar to the area of New Mexico state. 3002 images were collected between May and August of the year 2018 and 2019 from three different areas of New Mexico state (Las Cruces, Deming, and Los Lunas). Each category has equal values of 1501. The images with two class labels namely "Disease" and "Normal" were analysed with CNN using python 3.6 software.

Figure 1.1 Chile plant Normal

Figure 1.2 Chile plant with Disease

Prepossessing: Adequate model performance solely relies on whether the data is formatted into appropriate floating tensor points before feeding it into the network. The JPEG picture files was read in and decoded to RGB grids of pixels. Images were resized to 50 x 50 pixels and model normalization; optimization and predictions were performed on these downscaled plant images. Rescaled pixel values from 0 - 255 to 0 – 1. Keras package handles some prepossessing, batch size = 20.

Deep Neural Network (DNN) is an Artificial Neural Network (ANN) with multiple hidden layers between input and output layers. It can be supervised, partially supervised, or unsupervised. The supervised deep learning was considered in this study.

Loss Function: The robustness of any machine or deep learning model depends on how low the value of the loss function is. It is a very important part of artificial neural network modelling. It is used to measure how close is the predicted response $\hat{y}$ to actual labelled response y. the smaller the value of loss functions the better the model.

Gradient Descent: Gradient descent is essential to find the optimum value of θ that minimize the loss function ($J(\theta)$). The approach is the first order optimization process which is used to find local minima of an objective function. n Is the number of repetitions until convergence is reached. α learning rate. SGD requires having dataset in batches f_b, y_b. The mathematical representation of SGD is given as follows

Image augmentation technique was implemented by artificially expanding dataset. The parameter used were rotation, zoom, shear and preprocessing functions. In addition, augmented images were generated by custom function for contrast stretching, histogram equalization and adaptive histogram equalization. This research, the rotation range was set to be 40 degrees. The flip is horizontal with Width shift, height shift, shear range, and zoom range were all set to 0.2.

Results

Support vector Machine

The results for support vector machine with gamma 0.01 for linear kernel and list of C[10, 20, 30, 50, 100, 150, and 200] were given below. The results for support vector machine with (RBF) and list of kernels (c) [10, 20, 30, 50, 100, 150, and 200] were given below

Table 1: support vector machine (SVM) for RBF and for linear with gamma 0.01

(c)	Kernel	Misclassified samples	Accuracy (%)	Run Time
10	Linear	239	69	0:00:00.160
	RBF	244	69	0:00:00.365
20	Linear	235	70	0:00:00.156
	RBF	240	69	0:00:00.324
30	Linear	237	70	0:00:00.157
	RBF	237	70	0:00:00.322
50	Linear	238	70	0:00:00.187
	RBF	237	70	0:00:00.317
100	Linear	240	69	0:00:00.162
	RBF	**233**	**70**	**0:00:00.307**
150	Linear	240	69	0:00:00.191
	RBF	236	70	0:00:00.318
200	Linear	242	69	0:00:00.183
	RBF	235	70	0:00:00.325

Considering the SVM results from the table above, SVM model with C equals to C, Kernel = RBF, has the best performance with high accuracy of 70% and smallest misclassified sample of 233 in considerable low time of 307 split seconds

Table 2 Accuracy and F1-Score Results

		Training Acc	Validation Accuracy and F1-Score			
	Epoch	Total Acc (%)	Accuracy (%)	Class accuracy (%)		F1-Score (%)
				Disease	***Normal***	
CNN model	15	95.07	91.87	82.60	100	90.47
	30	97.54	98.02	97.30	99.73	98.51
CNN (Dropout)	15	93.36	92.70	97.30	90.35	91.01
	30	**97.63**	**98.19**	99.02	98.93	**98.97**
CNN(Augmentation)	15	95.28	95.50	95.59	94.91	95.13
	30	97.25	97.25	99.51	96.25	96.46

The table above shows the overall training accuracy, validation accuracy, class accuracy and F1-score. CNN with dropout regularization performed best with highest validation accuracy of 98.19%, and F1-Score of 98.97%

Training and validation Accuracy and loss

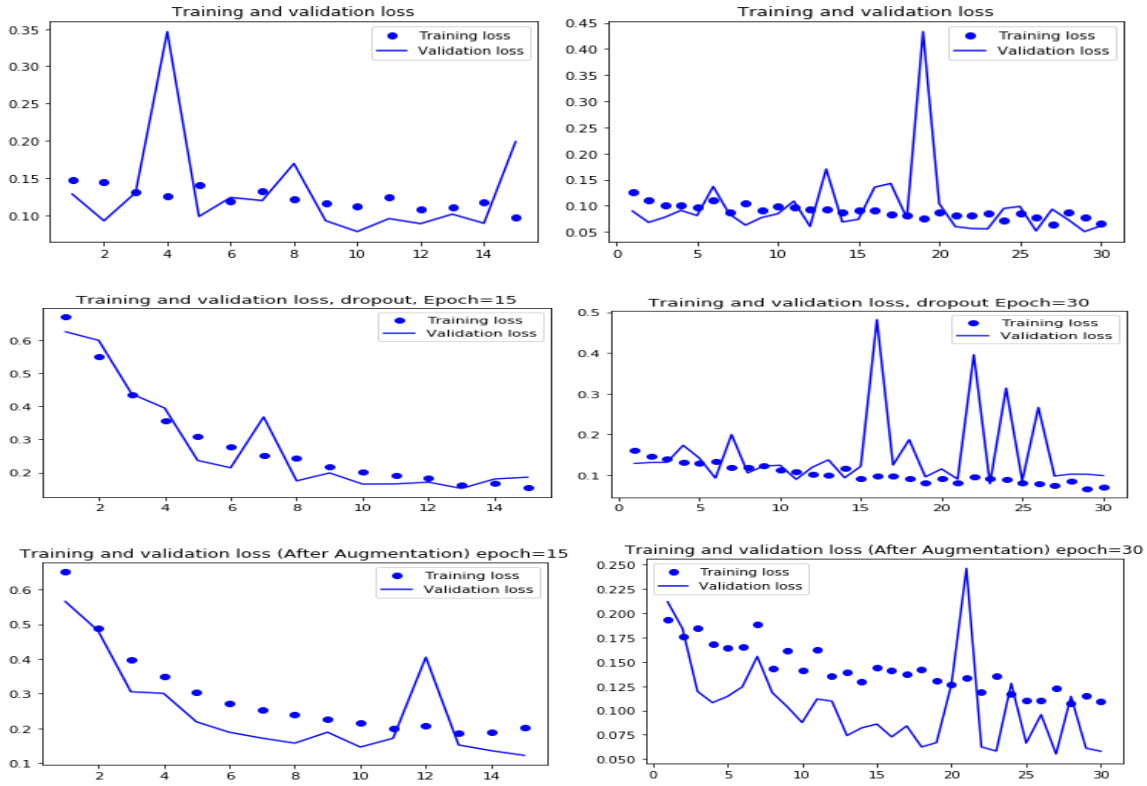

Figure 2 Training and Validation loss for CNN model

The results displayed in figure 1 are related to the training dataset containing performance base on original, Augmented, and Dropout regularized approaches.

Table 3 Test accuracy and F1-score results of CNN model for Chile plant image data

	Test **Accuracy and F1-Score**				
	Epoch	Accuracy (%)	Class accuracy (%)		F1-Score (%)
			Disease	***Normal***	
CNN model	15	70.00	42.00	98.00	58.33
	30	74.25	50.50	98.00	66.23
CNN (Dropout)	15	**92.25**	88.50	98.00	**91.95**
	30	87.50	77.00	97.50	86.03
CNN(Augmentation)	15	82.50	67.00	96.00	79.29
	30	81.00	64.50	98.00	77.25

The final test results give more confidence and reliability on the robustness of the data as dropout regularized model with epoch 15 proved to be the best among the considered model with: overall test accuracy = 92.25%, and F1-Score = 91.95%

Table 4: AUC values

	Epoch	**AUC**	**Epoch**	**AUC**
Regular CNN	15	0.70	30	0.75
CNN (Dropout)	15	**0.93**	30	0.88
CNN(Augmentation)	15	0.83	30	0.81

The AUC values in table 4 shows the respective success on the test set as convolunational neural network (CNN) with droupout regularization has highest Area Under the Curve (AUC).

Conclusion

The project developed a deep learning image processing technique with an achieved goal to build a robust model "CNN model" that can automatically detect Chile plant disease mostly caused by pathogenic disease. The technique involves acquiring, prepossessing, regularizing, and classification. The development of an automatic detection system from image processing techniques will help farmers in the area to identify Chile plant diseases early and take necessary control measures.

Among the considered models, (that is, Support Vector Machine 'SVM', and Deep learning.) It is generally true that complex models with many parameters tend to have high variance (although low bias effect), which in most cases results in overfitting, especially when datasets are limited. Data augmentation and Dropout regularization are the two powerful techniques considered in this project to mitigate these problems. Among the two techniques, Dropout regularization has the highest performance with overall test accuracy of 98.19% and class accuracy of 99.02% and 98.93% for Disease and Normal Chile plant, respectively (Table 2) and test accuracy of 92.25% from table 2. AUC also confirmed the performance of CNN with dropout regularized techniques with the value of 0.93.

We would like to extend our work further on more plant disease detection.

References

https://articles.mercola.com/herbs-spices/chili-peppers.aspx Chili Pepper Benefits, Uses, and Recipes - Mercola.com.

Chung, Chia-Lin, Kai-Jyun Huang, Szu-Yu Chen, Ming-Hsing Lai, Yu-Chia Chen, and Yan-Fu Kuo. 2016. 'Detecting Bakanae Disease in Rice Seedlings by Machine Vision'. *Computers and Electronics in Agriculture* 121:404–11.

Ebrahimi, M. A., M. H. Khoshtaghaza, Saeid Minaei, and B. Jamshidi. 2017. 'Vision-Based Pest Detection Based on SVM Classification Method'. *Computers and Electronics in Agriculture* 137:52–58.

Garg, Kshitiz and Shree K. Nayar. 2007. 'Vision and Rain'. *International Journal of Computer Vision* 75(1):3–27.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. 'Deep Learning'. *Nature* 521(7553):436.

Liakos, Konstantinos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. 2018. 'Machine Learning in Agriculture: A Review'. *Sensors* 18(8):2674.

Moshou, Dimitrios, Cédric Bravo, Jonathan West, Stijn Wahlen, Alastair McCartney, and Herman Ramon. 2004. 'Automatic Detection of 'Yellow Rust'in Wheat Using Reflectance Measurements and Neural

Networks'. *Computers and Electronics in Agriculture* 44(3):173–188.

Pantazi, Xanthoula Eirini, Alexandra A. Tamouridou, T. K. Alexandridis, Anastasia L. Lagopodi, G. Kontouris, and Dimitrios Moshou. 2017. 'Detection of Silybum Marianum Infection with Microbotryum Silybum Using VNIR Field Spectroscopy'. *Computers and Electronics in Agriculture* 137:130–137.

Salakhutdinov, Ruslan and Geoffrey Hinton. 2009. 'Deep Boltzmann Machines'. Pp. 448–455 in *Artificial intelligence and statistics*.

Savary, Serge, Andrea Ficke, Jean-Noël Aubertot, and Clayton Hollier. 2012. *Crop Losses Due to Diseases and Their Implications for Global Food Production Losses and Food Security*. Springer.

Wolff, Franziska. 2004. 'Legal Factors Driving Agrobiodiversity Loss'. *Environmental Law Network International* 1:2004.