ISSN 2278-3091

Volume 9, No.1, January – February 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse67912020.pdf

https://doi.org/10.30534/ijatcse/2020/67912020

Machine Learning Approach towards Tomato Leaf Disease Classification

H. D. Gadade[1], Dr. D.K.Kirange[2]

[1]Government College of Engineering, Jalgaon, India, harishgadade2012@gmail.com

[2]J.T.Mahajan College of Engineering, Faizpur, India, dkirange@rediffmail.com

ABSTRACT

India in an agricultural country and the detection of diseases in first stage is very important to increase the crop yield. The bacterial spot, late blight, septoria leaf spot and yellow curved leaf diseases affect the crop quality of tomatoes. In this paper, to detect symptoms of disease, we have developed a module that classifies the plant leaf disease automatically. This paper presents a performance measure for different feature extraction techniques for tomato leaf disease detection including GLCM, Gabor and SURF and classification techniques including decision trees, SVM, KNN and Naïve Bayes. The dataset contains 500 images of tomato leaves with seven symptoms of diseases. We have modeled a system for automatic feature extraction and classification. We have evaluated the performance of the system using different performance measures to conclude with appropriate features set and classification technique for tomato leaf disease classification. The experimental results validate that Gabor features effectively recognizes different types of tomato leaf diseases. Accuracy of SVM is better as compared to other classification techniques but the execution time is more.

Key words: Decision Trees, Gabor GLCM, KNN, Naïve Bayes, SURF, SVM.

1. INTRODUCTION

Indian economy is depends on agriculture and most of the peoples depends on agricultural farming. Infected plants and crops cause the reduction in qualitative and quantitative yield. Therefore, challenging task is to identify plant leaf disease very accurately. Disease attacking parts are leaves, stems and fruits. Accurate quantification of diseases with necked eyes is very difficult. Hence understanding of specific and accurate image pattern is demanding now a day. In biological science, vast amount of images getting generated with single experiments and these images further can be used for classification. Hence biologists need to analyze and extract the specific contents for further classification. And here the role of image processing plays an important role. The diseases are caused because of environment change and it destroys the significant amount of crop yield. The common diseases are like fungi, bacteria and viruses, and due to adverse environmental conditions. Therefore, in an initial stage, diagnosis of disease is most import task. Continuous inspection of crop by the farmer with expert system is essential. Therefore, automatic and less expensive detection of disease is vital in farming. Most of the tomato leafs are attacked with fungi, bacteria and viruses. Sample examples of common tomato leaf disease are shown in Figure 1.

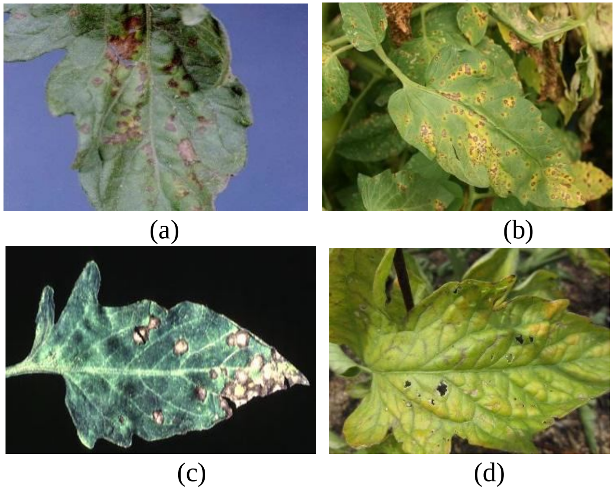
(a) (b) (c) (d)

Figure 1: Samples of tomato plant diseases; (a) Early Blight; (b) Septoria Leaf Spot; (c) Bacterial Spot (d) Iron Chlirosis

This research paper aims at following objectives

- To identify deficiency of tomato plant by analyzing its leaf efficiently.
- By identifying deficiency of tomato leaf, we can obtain healthy products (tomatoes).
- To predict occurrence of disease accurately based on analyzing deficiency symptoms.

The rest of the paper is organized as follows: The dataset used for tomato leaf disease classification is discussed in section 2, related work is presented in the section 3, followed by the proposed method in the section 4, while experimental set-up and results are discussed in the section 5. Finally, we draw our conclusion in the section 6.

2. PLANTVILLAGE IMAGE DATASET FOR TOMATO LEAF DISEASE CLASSIFICATION

We have collected vast amount of data from well-known PlantVillage Datasets[1] and it has around 50,000 images of 14 crop species and 26 diseases. We are working on 9,000 images of tomato leaves. Our datasets has 7 (seven) types of diseased images [1]:

- Class (0): Bacterial Spot.
- Class (1): Early Blight.
- Class (2): Healthy.
- Class (3): Septorial Leaf Spot.
- Class (4): Leaf mold.
- Class (5): Yellow Leaf Curl Virus.
- Class (6) LateBlight
- Class (7) Tomato Mosaic

3. RELATED WORK

In paper [2] four steps have been demonstrated. In first step, images have been gathered from different parts of country for training and testing purpose. Gaussian filter is applied in second part to remove noise and to get green color component, thresholding is done. Segmentation is done by means of K-means clustering and HSV image is taken from RGB color to extract color. The paper [3], jute plant disease is presented by means of image processing. Once image is captured and realized to match the size of the image which is to be stored in the database. The obtained image is enhanced in quality and noise is to be removed. Region of interest is extracted by converting image from RGB to HSV. This approach proposed can significantly support detecting stem oriented diseases for jute plant. In paper [5], detection of unhealthy plant leaves is done using image processing and genetic algorithm with Ardunio system. Paper [6] includes tomato disease detection using computer vision in which , depending on threshold value, a gray scale image is turned into binary image. The methodology for cucumber disease detection is presented in paper [7]. The methodology includes traditional steps like image acquisition, image preprocessing, feature extraction with Gray level co-occurrence matrix (GLCM) and finally classified using SVM. In paper [8], enhancement techniques like histogram equalization and contrast adjustment said to be used to improve the image quality. In paper [9], to detects diseses the agriculture product, popular methods have been utilizes machine learning, image processing and classification based approaches. In paper [10] image processing technique are used to detect the citrus leaf disease. This system includes: Image preprocessing, segmentation of the leaf using K-means clustering to determine the diseased areas, feature extraction and classification of disease. Uses Gray-Level Co-Occurrence matrix (GLCM) for feature extraction and classification is done using support vector machine (SVM)[2]. Paper [11] presents classification and detection techniques that can be used for plant leaf disease classification. Here preprocess is done before feature extraction. RGB images are converted into white and then converted into grey level image to extract the image of vein from each leaf. Then basic Morphological functions are applied on the image. Then the image is converted into binary image. After that if binary pixel value is 0 it's converted to corresponding RGB image value. Finally by using Pearson correlation and Dominating feature set and Naïve Bayesian classifier disease is detected.

4. TOMATO LEAF DISEASE CLASSIFICATION

The system contains five major modules: tomato leaf disease input database, preprocessing/ Noise removal, feature extraction, classifier & recognized output as illustrated in figure 2. Overall, the system is based on preprocessing/ noise removing mechanism of leaf disease images, extracting some of features which contain information about textural features of the image & taking appropriate pattern recognition model to identify the type of tomato leaf image disease

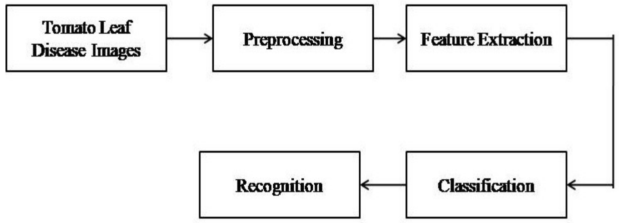

Figure 2. Structure of Tomato Leaf Image Disease Classification

Various steps for classification of tomato disease from disease leaf images as follows:

1. Database Gathering and Preprocessing:Obtaining the tomato leaf disease images and Normal tomato leaf images Data from Plant Village Dataset
2. Preprocessing and noise removal of tomato leaf images using Median Filter.
3. Feature Extraction : GLCM, Gabor and SURF features
4. Design and development of the system for classification of tomato leaf images as normal or diseased containing 7 types of tomato diseases.
5. Evaluate the performance of different classifiers
   - SVM
   - KNN
   - Naïve Bayes
   - Decision Trees

4.1 Gabor Feature Extraction

Gabor Filters is an efficient and very effective extraction technique to extract texture feature and their analysis. it is subjected to work on the frequency patterns of a particular location or region of interest and then matching is done. A Gabor filter is essentially a sinusoidal signal with a given frequency and orientation, modulated by a Gaussian. We have used it for edge and object detection, coding, color or pattern gradient and image representation etc.

4.2 SURF Features

Speeded up robust features (SURF) is a local feature detector and descriptor. Image registration, object recognition is the primary tasks of SURF. It can be used for tasks such as object recognition, image registration, classification, or 3D reconstruction. Fast computation of operators using box filter con be done by means of SURF that enables real time applications like tracking and visual perception.

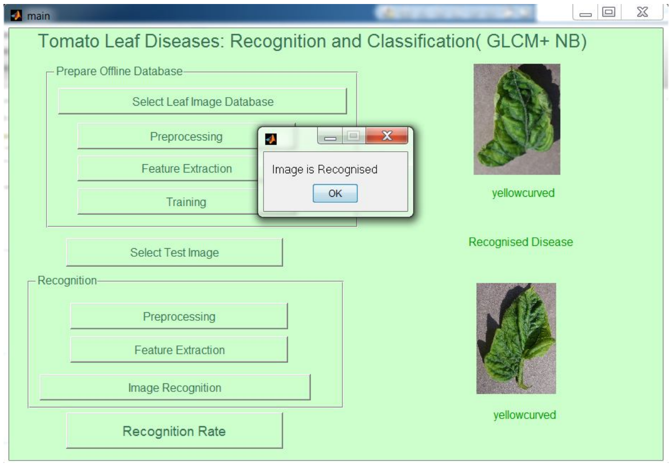

Image is Recognised

Figure 3. Matlab Based GUI for Tomato Leaf Disease Classification

4.3 Statistical Features

$$v = gcp(glcm, \omega) \qquad (1)$$

Here v is an array of statistics for each $glcm$, ω is the properties and v Calculates the statistics specified in properties from the gray-level co-occurrence matrix $glcm$. gcp Normalizes the gray-level co-occurrence matrix (GLCM) that mean to sum of its elements is equal to 1. Different statistical features considered here are contrast, correlation, Energy and Homogeneity are the different statistical features considered in this method.

4.4 Support Vector Machine

Supervised learning model called support-vector machine (SVM also called as support vector networks) analyze the data used for classification and regression analysis [2]. Given a set of training examples, each marked as belonging to one or the other of two categories, a model which is said to be build by SVM training algorithm that assigns a new examples to one or another category, preparing it as a non-probabilistic binary linear classifier[13].

Linear classification and non-linear classifications can be performed by SVM efficiently by means of kernel trick. For unlabelled data, supervised learning is not possible and to find natural clustering of the data to groups, and then map new data to these formed groups, an unsupervised learning approach is required[2].

4.5 KNN Classification

A non-parametric method called k-nearest neighbor algorithm(KNN) used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. Class membership is the output in KNN methods. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor. The output is the valuable in KNN regression and this value is the average of the values of k nearest neighbors.

4.6 Naïve Bayes(NB) Classification

Based on Baye's theorem, Naive Bayes (NB) classifiers is a collection of classification algorithms. All algorithms from its famiy, shares common principles i.e. every pair of features being classified is independent of each other. The Naïve Bayes classifier assumes independence between predictor variables conditional on the response, and a Gaussian distribution of numeric predictors with mean and standard deviation computed from the training dataset[13].

4.6 Decision Tree

Decision tree classifiers are successfully used in diversified areas[13]. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (s), each belonging to one of the classes $C1, C2, ..., Ck$ is as follows:

Step 1: If all the objects in **S** belong to the same class, for example C_i, the decision tree for s consists of a leaf labeled with this class.
Step 2: Otherwise, let T be some test with possible outcomes $O1, O2, ..., On$

Each object in S has one outcome for T so the test partitions S into subsets $S1, S2, ...Sn$ where each object in S_i has outcome O_i for T. T becomes the root of the decision tree and for each outcome O_i we build a subsidiary decision tree by invoking the same procedure recursively on the set S_i.

5. RESULT ANALYSIS

To classify efficiently a tomato leaf diseases, a Matlab-based GUI-driven tool is developed. Figure 3 shows graphical user interface(GUI) developed for proposed algorithm before execution. GUI for this software is divided into number of subgroups according to their functionality.

5.1 Database Selection and Preprocessing

Tomato leaf disease images training database is selected. Noise is removed by means of median filter in preprocessing.

5.2 Features Extraction

From the preprocessed training images Gabor, SURF and statistical features are extracted. Features matrix is constructed.

5.3 Classification

Different classifiers including SVM, KNN, Naïve Bayes and Decision Trees are trained with various features for tomato leaf disease classification. This module deals with tomato leaf disease detection and classification. The performances of different classifiers have been evaluated by considering different number of training images. Four parameters are used for evaluating performance of the algorithm. Those are accuracy, precision, recall and F measure. These parameters are defined using four measures; True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

True Positive: DR detection coincides with actual labelled data
True Negative: both classifier and actually labelled absence of DR
False Positive: system labels a healthy case as an DR one
False Negative: system labels DR image as healthy

Accuracy: Accuracy is the ratio of number of correctly classified cases[12][14], and is given by

$$Accuracy = \frac{(TP + TN)}{N} \qquad (2)$$

Total numbers of cases are N

Precision: Precision is the fraction of retrieved images that are relevant to the query. Precision takes all retrieved images into account, but it can also be evaluated at a given cut-off rank, considering only the results returned by the system[12][14].
Precision is defined as

$$\mathrm{Pr}\,ecision = \frac{TP}{(TP + FP)} \qquad (3)$$

Recall: The Fraction of relevant images said to be retrieved successfully is called recall. In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query. To achieve recall of 100%, all documents returns in response to any query[12],[14].
Recall is defined as

$$\mathrm{Re}\,call = \frac{TP}{(TP + FN)} \qquad (4)$$

The weighted average of Precision and Recall is $F1$ Score. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but $F1$ is usually more useful than accuracy, especially if you have an uneven class distribution. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In our case, $F1$ score[12] is 0.701.

$$F1Score = 2 * \frac{(\mathrm{Re}\,call * \mathrm{Pr}\,ecision)}{(\mathrm{Re}\,call + \mathrm{Pr}\,ecision)} \qquad (5)$$

Table 1: Performance measure with decision tree

Decision Trees				
	'Accuracy'	'Precision'	'Recall'	'F Measure'
GLCM	0.6497	0.1715	0.722	0.2772
Gabor	**0.6759**	**0.2047**	**0.861**	**0.3307**
SRF	0.5073	0.1033	0.5593	0.1743

Table 2.:Performance measure with svm

SVM				
	'Accuracy'	'Precision'	'Recall'	'F Measure'
GLCM	0.4095	0.1306	0.9458	0.2295
Gabor	**0.7339**	**0.2525**	**0.9492**	**0.3989**
SRF	0.326	0.1084	0.8644	0.1926

Table 3: Performance measure with knn

KNN				
	'Accuracy'	'Precision'	'Recall'	'F Measure'
GLCM	0.633	0.1701	0.7593	0.2779
Gabor	**0.732**	**0.2555**	**0.9831**	**0.4056**
SRF	0.5485	0.1462	0.7966	0.2471

Table 4: Performance measure with naïve bayes

NB				
	'Accuracy'	'Precision'	'Recall'	'F Measure'
GLCM	0.6589	0.1919	0.8305	0.3117
Gabor	**0.675**	**0.2187**	**0.9695**	**0.3568**
SRF	0.3493	0.0876	0.6373	0.1541

As depicted in tables I to IV, Gabor features with all classifiers give promising result for tomato leaf disease classification. Table V shows the performance evaluation of the algorithms in terms of execution time. The SVM classifier requires more execution time for training and evaluation.

Table 5:. Performance Evaluation of algorithms

	Decision Trees	SVM	KNN	NB
GLCM	1.0015	20.5218	3.8556	0.7223
Gabor	44.9283	142.5488	41.7596	29.8614
SRF	2.8596	526.2486	3.7722	1.6257

Table 6 shows the performance measure of different classifiers using Gabor features. The Gabor features with SVM classification gives better performance but the execution time required for SVM is more.

Figure 4 shows the performance measure of different classifiers using Gabor features for tomato leaf disease classification.

Table 6:.Performance measure with gabor features

	Gabor		
	'Accuracy'	'F Measure'	'Time Elapsed'
Decision Trees	0.67	0.33	44.9283
SVM	**0.73**	0.39	142.5488
KNN	0.73	0.40	**41.7596**
NB	0.67	0.35	29.8614

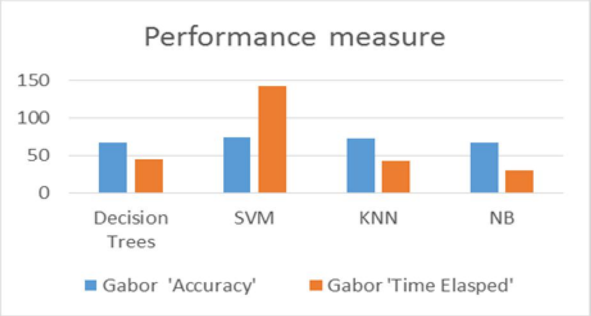

Figure 4: Performance Measure Using Gabor Features

As depicted in Fig.4 the accuracy of SVM classification is better but the execution time is more. So for tomato leaf diseases classification, KNN classification is better with less execution time and better performance.

6. CONCLUSION

In this paper, KNN classification framework with Gabor features is used for tomato leaf disease classification. Different features like SURF, Statistical and Gabor are used. The different classifiers including SVM, KNN, Naïve Bayes and decision trees are trained to carry out the final classification. Main focus of this study is to preprocess the tomato leaf images for noise removal. After preprocessing and features extraction, classification of the selected seven different tomato leaf diseases is performed. For PlantVillage data KNN with Gabor features gives better accuracy in terms of precision, recall and F measure. KNN also needs lower execution time as compared to SVM. The experimental results have demonstrated the effectiveness of our proposed algorithm to be good enough to be employed in real time applications. KNN enough accurate by means of classification result. The method proposed in the paper could also use for plant disease image recognition and classification. There are more sophisticated techniques are available for classification like Adaptive neuro fuzzy, Neural Networks, Genetic algorithm. etc. for image classification. These techniques can also use for plant image recognition and classification.

ACKNOWLEDGEMENT

I would like to be very thankful to my guide Dr. D.K. Kirange for continuous guidance, support and motivation toward research. I would like to thank my research laboratory at J.T. Mahajan College of Engineering, Faizpur for providing infrastructure to carry out this research. I would like to thank Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon for giving me a chance to carry out this research. I also thank to TEQIP-III, government college of Engineering, Jalgaon for providing me a financial and research facilities. I would also be thankful to Dr. R. D. Kokate principal, Government College of Engineering, Jalgaon for motivation and encouragement. I would also be thankful to Dr. R. D. Patil, Principal, J.T.Mahajan College of Engineering, Faizpur for guidance and support.

REFERENCES

1. https://www.kaggle.com/emmarex/plantdisease, was retrieved 20/12/2018.
2. Pranjali B. Padol, Prof. AnjilA.Yadav, **"SVM Classifier Based Grape Leaf Disease Detection"** 2016 Conference on Advances in Signal Processing(CAPS) Cummins college of Engineering for Women, Pune. , pp 175-179, IEEE, ISBN 978-1-5090-0849-0, DOI 10.1109/CASP.2016.7746160, 2016.
3. Zarreen Naowal Reza, Faiza Nuzhat, Nuzhat Ashraf Mahsa, Md. Haider Ali, **“Detecting jute plant disease using image processing and machine learning”**, 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), IEEE, ISBN 978-1-5090-2906-8, DOI 10.1109/CEEICT.2016.7873147, 2016
4. Tejoindhi M.R, Nanjesh B.R, Jagadeesh Gujanuru Math, Ashwin Geet D'sa **"Plant Disease Analysis Using Histogram Matching Based On Bhattacharya's Distance Calculation"** International Conference on Electrical, Electroniocs and Optimization Techniques(ICEEOT), ISBN 978-1-4673-9939-5 DOI: 10.1109/ ICEEOT. 2016.7754943, 2016.
5. Arya M S, Anjali K, Mrs.Divya Unni, **“Detection of unhealthy plant leaves using image processing and genetic algorithm with Arduino”**, 2018 International Conference on Power, Signals, Control and Computation (EPSCICON), IEEE, ISBN 978-1-5386-4208-5, DOI 10.1109/EPSCICON.2018.8379584, 2018
6. Tanvimehera, vinaykumar,pragyagupta **"Maturity and disease detection in tomato using computer vision"** 2016 Fourth international conference on parallel, distributed and grid computing(PDGC), pp 399-403, IEEE, ISBN 978-1-5090-3669-1, DOI 10.1109/PDGC.2016.7913228, 2016
7. Ms.Pooja pawer, Dr.varshaTukar, prof.parvin patil **"Cucumber Disease detection using artificial neural network"**, International Conference on Inventive

Computation Technologies (ICICT), IEEE, ISBN 978-1-5090-1285-5, DOI 10.1109/INVENTIVE.2016.7830151, 2016

8. R.P.Narmadha , G.Arulvadivu ,**“Detection and measurement of paddy leaf disease symptoms using image processing”**, 2017 International Conference on Computer Communication and Informatics (ICCCI -2017), IEEE, ISBN: 978-1-4673-8855-9, DOI 10.1109/ICCCI.2017.8117730, Jan. 05 – 07, 2017, Coimbatore, INDIA
9. Mukesh Kumar Tripathi, Dr.Dhananjay, D.Maktedar, **''Recent Machine Learning Based Approaches for Disease Detection and Classification of Agricultural Products''** International Conference on Electrical, Electronics and Optimization Techniques (ICEEOT),IEEE, ISBN: 978-1-5090-3291-4, DOI 10.1109/ ICCUBEA.2016.7860043, 2016.
10. R.Meena Prakash, G.P.Saraswathy, G.Ramalakshmi, **“Detection of leaf diseases and classification using digital image processing”**, 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS),IEEE, ISBN: 978-1-5090-3294-5, DOI 10.1109/ICIIECS.2017.8275915, 2017
11. Dhiman Mondal, Dipak Kumar Kole, Aruna Chakraborty, D. Dutta Majumder, **"Detection and Classification Technique of Yellow Vein Mosaic Virus Disease in Okra Leaf Images using Leaf Vein Extraction and Naive Bayesian Classifier”**,2015, International Conference on Soft Computing Techniques and Implementations- (ICSCTI), ISBN: 978-1-4673-6792-9, DOI 10.1109/ICSCTI.2015.7489626, 2015.
12. S Jafar Ali Ibrahim, Dr.M.Thangamani, **“Innovative Drug and Disease Prediction with Dimensionality Reduction and Intelligence Based Random Walk Methods”**, International Journal of Advanced Trends in Computer Science and Engineering, ISSN 2278-3091, Volume 8, No.4, July – August 2019. https://doi.org/10.30534/ijatcse/2019/93842019
13. Munya A. Arasi, Sangita babu, **“Survey of Machine Learning Techniques in Medical Imaging”**, International Journal of Advanced Trends in Computer Science and Engineering, ISSN 2278-3091, Volume 8, No.5, September-October 2019. https://doi.org/10.30534/ijatcse/2019/39852019
14. Rein Rachman Putra, Monika Evelin Johan, Emil Robert Kaburuan, **“A Naïve Bayes Sentiment Analysis for Fintech Mobile Application User Review in Indonesia”**, International Journal of Advanced Trends in Computer Science and Engineering, ISSN 2278-3091, Volume 8, No.5, September-October 2019.