# Tomato leaves diseases detection approach based on Support Vector Machines

**Conference Paper** · December 2015

DOI: 10.1109/ICENCO.2015.7416356



**4 authors:**

Mona Abdelbaset Sadek Ali
Benha University


Usama Mokhtar
Institute of Statistical Studies and Research


Aboul Ella Hassanien
Cairo University


Hesham A. Hefny
Faculty of Graduate Studies for Statistical Research (FGSSR)

# Tomato leaves diseases detection approach based on support vector machines

Usama Mokhtar*, Mona A. S. Ali†, Aboul Ella Hassenian‡, Hesham Hefny§
*Inst. of Stat. Studies and Res., Cairo University, Egypt
†Faculty of Computers and Information, Minia University, Egypt
‡Faculty of Computers and Information, Cairo University, Egypt
Scientific Research Group in Egypt (SRGE), http://www.egyptscience.net
§Inst. of Stat. Studies and Res., Cairo University, Egypt

***Abstract*—The study described in this paper consists of a method that applies gabor wavelet transform technique to extract relevant features related to image of tomato leaf in conjunction with using Support Vector Machines (SVMs) with alternate kernel functions in order to detect and identify type of disease that infects tomato plant. Initially, we collected real samples of diseased tomato leaves, next we isolated each leaf in single image, wavelet based feature technique has been employed to identify an optimal feature subset. Finally, a support vector machine classier with different kernel functions including Cauchy kernel, Invmult Kernel and Laplacian Kernel was employed to evaluate the ability of this approach to detect and identify where tomato leaf infected with Powdery mildew or early blight.**

**To evaluate the performance of presented approach, we present tests on dataset consisted of 100 images for each type of tomato diseases. Extensive experimental results demonstrate that the proposed approach provides excellent annotation with accuracy 99.5 %. Efficient result obtained from the proposed approach can lead to tighter connection between agriculture specialists and computer system, yielding more effective and reliable results.**

***Index Terms*—image processing, K-Mean clustering algorithm, Geometric features, support vector machine (SVM)**

## I. Introduction

Tomatoes are one of the most widely cultivated food crops throughout the world due to its high nutritive value. It contains a lot of vitamins and nutrients such that vitamin C. It occupies the fourth level between word vegetables. Egypt is one of the famous countries that interested in tomatoes cultivation. It ranked fifth among leader countries in the world [1]. Although, the naked eye observation of experts is the main approach adopted in practice for detection and identification of plant diseases, it requires continuous monitoring of experts, which might be expensive and difficult especially in large farms. so, It is necessary to help farmers in automatically detect symptoms of disease as soon as they appear by analyzing the digital images. Aiming at minimizing major production and economic losses, ensuring both quality and quantity of agricultural products and minimizing agrochemicals use, computer based applications have been developed and revealed high efficacy, many of them focusing on the identification of diseases through foliage symptoms in various cultivars, such as: wheat [2], cotton [3], rice [4] and apple [5]. In this paper, a combination of various machine intelligence techniques has been used to detect diseased tomato leaf for the purpose of diagnosing pathogen types that infect tomato. The proposed SVM-based tomato diseases detection approach utilizes some preprocessing techniques such that image cropping and image filtering for improving the quality of leaf images. Gabor wavelet transform based feature technique was employed in order to identify significant feature subset that describe the diseased tomato leaves. Finally, support vector machine with different kernel functions including Cauchy kernel, Invmult Kernel and Laplacian Kernel functions was then performed for the actual diagnosis. For tuning and parameters optimization of kernels functions, we used Grid search along with N-cross validation techniques during training and testing phase. This paper is organized as follow: section 1 is introduction to the paper. Proposed system is introduced in section 2. Result of some experiments is introduced in section 3. The last section summarizes our contribution and focuses on possible directions of further research.

## II. Preliminaries

Due to space limitations, a brief explanation of basic technologies used in this paper including Gabor wavelet transformation and The support vector machine is introduced in the following subsection.

### A. Gabor wavelet transformation

Wavelets are a class of functions used to localize a given signal in both space and scaling domains [12]. The fundamental idea behind wavelets is to analyze the signal at different scales or resolutions, which is called multi-resolution [6]. Among various wavelet bases, gabor functions provide the optimal resolution in both the time (spatial) and frequency domains. Gabor transform and gabor wavelet are widely applied to image processing, computer vision and pattern recognition [7]. This function can provide accurate time-frequency location governed by the Uncertainty Principle [8], [9]. For these reasons, gabor wavelet transform is used in this research as effective feature extraction technique.

### B. The support vector machine

Support vector machine (SVM) is a classication technique originally developed by Vapnik [10]. It was recently used in many applications including pattern recognition, bioinformatics, cancer diagnosis [12]. It tries to find the tradeoff between minimizing the training set error and maximizing the margin, in order to achieve the best generalization ability and remains resistant to over fitting [11]. SVM works by mapping data to a high-dimensional feature space, so that data points can be categorized, even when the data are not otherwise linearly separable. SVM uses a nonlinear mapping to transform the original training data into a higher dimension , then it searches to nd a linear optimal hyperplane, so that the margin of separation between the two classes is maximized [13].

## III. The proposed SVM-based tomato diseases detection approach

The proposed SVM-based tomato diseases detection approach is comprised of the following four fundamental building phases: (1) **Image acquisition phase:** The proposed SVM-based tomato diseases detection approach begins with acquiring and collecting digital images from suitable environment.

(2) **Pre-processing phase:** To improve the quality of the images and to make the feature extraction phase more reliable prepare, some image-processing techniques are applied including leaf image isolation, image resizing and background removing. (3) **Feature extraction phase:** 402 texture-based features have been extracted, and represented in a database as vector values and (4) **Classification phase:** The last phase is the classification and prediction of new objects, it is dependent on the SVM classifier. These four phases are described in detail in the following subsections and depicted in figure 1.

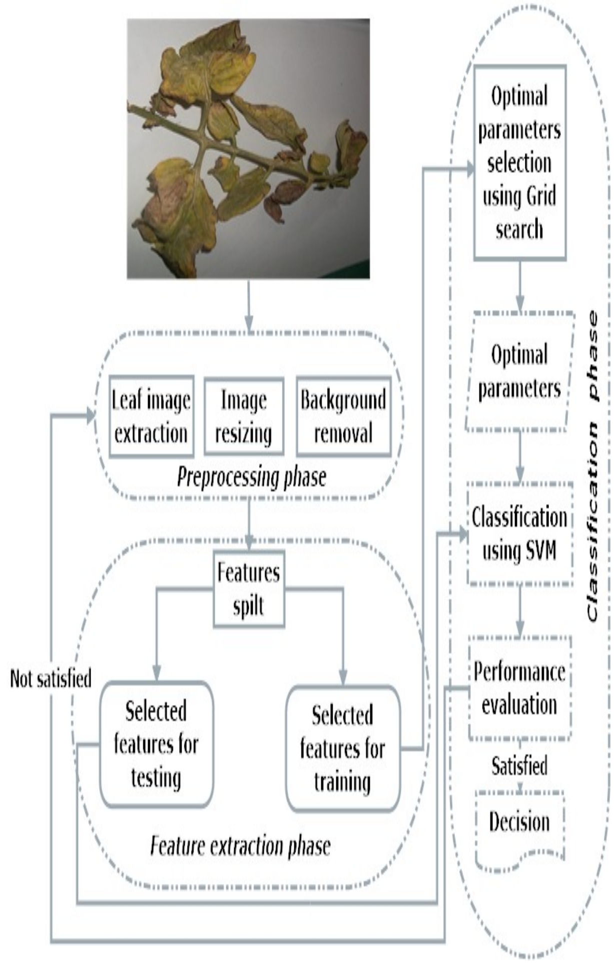



Fig. 1: Layout structure of the proposed SVM-based tomato diseases detection approach

### A. Image Acquisition phase

The first phase of the proposed SVM-based tomato diseases detection approach is the image acquisition stage. This phase plays an important role in any image classification system. We must select these images carefully to achieve the intended task. The datasets used for experiments were constructed based on real sample images of tomato leaves infected with two types of tomato diseases including Powdery mildew and early blight. this dataset were collected from different farms in bani seef city at February, at temperature between 16 and 20 degree. Since, some of these images contain more than one tomato leaf, therefore we addressed this issue in the preprocessing phase. For more details, see the next subsection. At the end of this phase, dataset consisted of 200 images of infected tomato leaves 100 for each virus type have been extracted.

### B. Pre-processing phase

Since the main objective of image pre-processing is to enhance, smoothness, remove noise that caused by defects of camera flash or hight lights and to increase the efficiency of the classification and prediction process, a pre-processing phase should be considered to enhance the quality of the input tomato leaves images and to make the feature extraction.

**leaf images isolation and extraction:** As we mentioned before, one of the main contributions of this paper is to employee the proposed technique on real dataset that we manually have been collected from several open farms, so most of acquired images have more than one leaf image and may be other things. since, we deal with each leaf separately, firstly, so we must isolate and extract every leaf in single image. In order to achieve that, these image have been manually cropped to extract a single leaf that saved in jpeg format.

**Image Resizing:** Since the input images may be of different size, Initially, captured images were resized to 512 x 512 resolution, so as to utilize the storage capacity or to reduce the computational time in the later processing.

**background removal:** After leaf image extraction, some small parts and shadow might be remained that would disturb the feature extraction phase. So the background of each image is removed using background subtraction technique with some morphological operations.

### C. Feature extraction phase based on gabor wavelet transform

This section illustrates the technique of how to get the feature vectors that used as input to classification phase. In order to recognise the disease category, we measured several number of features in the diseased image, to be later use for classification. The purpose of this phase is to reduce the original dataset by measuring certain features or properties of each image including texture and color. In this paper, Gabor transform is typically used to describe the textural pattern of diseased tomato leaves. In order to extract feature vectors, we divided the image that we obtained as output of the pre-processing phase into $K$ non overlapping regions of size $l \times l$. As we mention before a bank of $S \times L$ Gabor filters are employed for extracting texture features of each leaf. The average output for every $i$th filter in the region $k$ is obtained. For every filter $i$ in the filter bank, a cost function $J(i)$ is calculated. For more detail reader may be refere to [14]

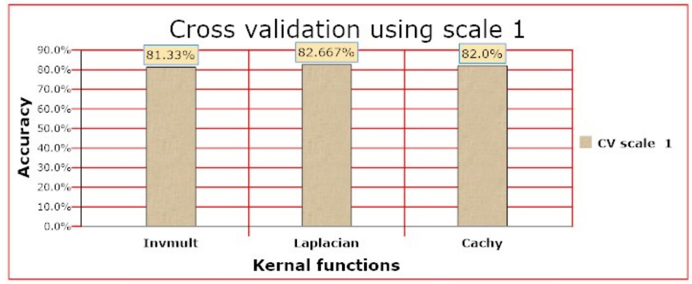
### *D. Classification phase based on SVM*

The final phase of this work is a classification phase. In this paper SVM has been employed to test the performance of the proposed SVM-based tomato diseases detection approach in identifying category of diseases that infect the tomato leaves. The inputs of this phase are training dataset feature vectors and their corresponding classes, whereas the outputs are the decision that determines type of disease that infected the tomato leaf. Since kernel function selection is one of the critical issues that affect the classification results, another contribution of this research is that SVM was trained and tested using the different kernel functions including Cauchy kernel, Invmult Kernel and Laplacian Kernel to compare between these different results.

To evaluate the performance of a classification system, cross-validation; a common method to deal with small training sets in machine learning [16] is commonly used. Cross-validation means to randomly divide the input examples into a number of equally sized subsets, and to train the classifier multiple times, each time on all but one of the subsets. After each training procedure, the excluded subset is used for validation [17].

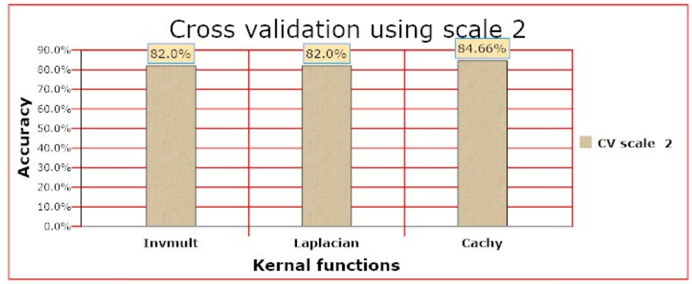
For the process of result evaluation, $k$ -fold cross-validation procedure has been employed. Firstly, dataset is divided into equally (or nearly equally) $K$-subsets. Then, the cross-validation process is performed using $(K-1)$ subsets for training, and 1 subset for testing. This is repeated $k$ times to use all possible combinations of training and validation sets. Hence we take the average of performance of $N$ runs. The algorithm performance measure can be calculated as the average of the performances of 50 runs.

Selection of optimal parameters of kernel functions is one of critical issues to develop the good SVM model. The best set of hyperparameters for the dataset at hand are usually approximated by a so-called grid search [17]. This approach is performed by training and validating SVM classifier multiple times using inner $k$ fold cross validation scheme within each training set. At the end of this step, a set of hyperparameters with the best validation results are selected.



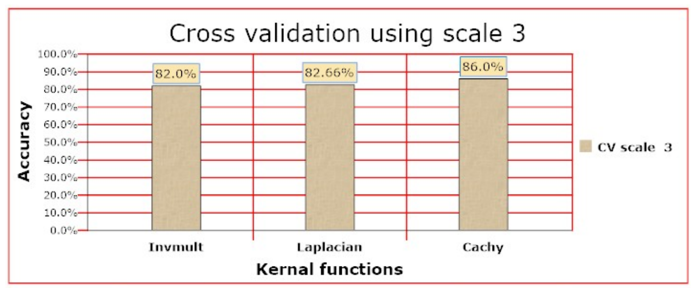

| | Invmult | Laplacian | Cachy |
|---|---|---|---|
| CV scale 1 | 81.33% | 82.667% | 82.0% |

(a) Cross validation accuracy using big scale



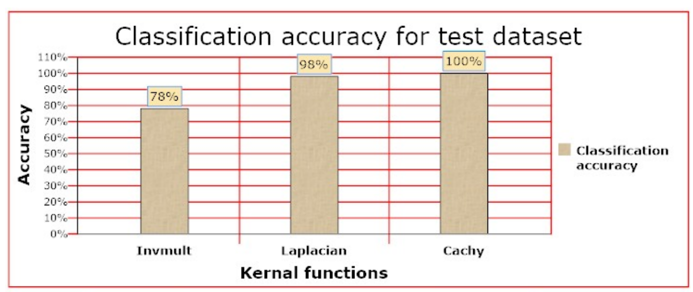

| | Invmult | Laplacian | Cachy |
|---|---|---|---|
| CV scale 2 | 82.0% | 82.0% | 84.66% |

(b) Cross validation accuracy using medium scale




| | Invmult | Laplacian | Cachy |
|---|---|---|---|
| CV scale 3 | 82.0% | 82.66% | 86.0% |

(c) Cross validation accuracy using small scale



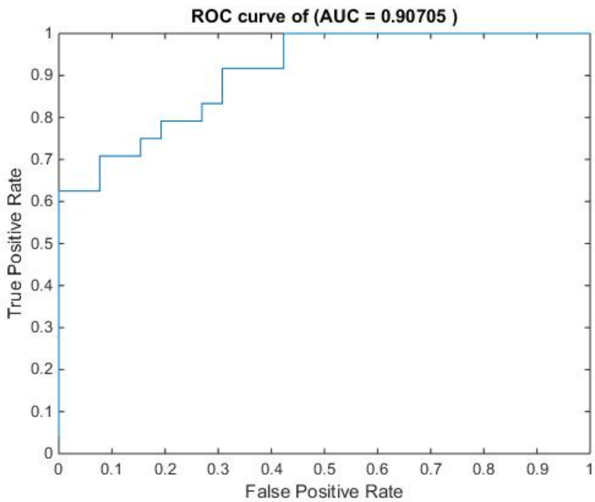

| | Invmult | Laplacian | Cachy |
|---|---|---|---|
| Classification accuracy | 78% | 98% | 100% |

(d) Classification accuracy using test dataset

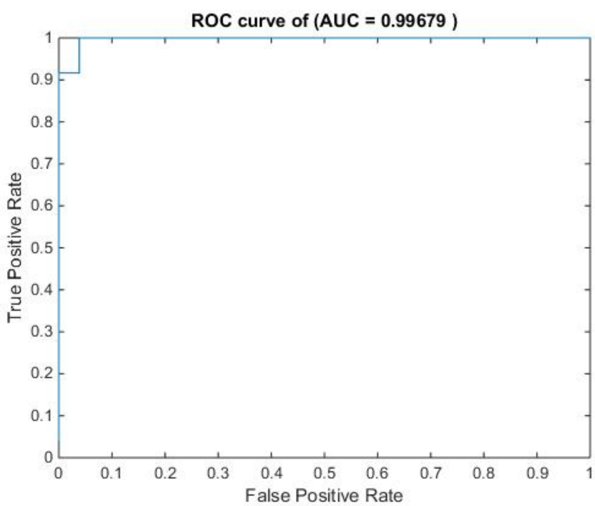
Fig. 2: Classification accuracy in training and test phase employing SVM with different kernel functions

## IV. Experimental results

Simulation experiments in this article are done on a laptop with Intel(R) Core (TM) i7 CPU Q820@1.73 GHZ and 8 GB memory. The proposed SVM-based tomato diseases detection approach is designed with Matlab running on Windows 7.




(a) ROC curve for SVM using invmult kernel




(b) ROC curve for SVM using Laplacian kernel

Fig. 3: ROC curve and (AUC) for the SVM with different kernel functions

The datasets used for experiments were constructed based on real sample images of diseased tomato leaves. These dataset were collected from different farms in bani seef city in February at temperature between 16 and 20 degree. Directly met the farmers and get the suggestion from them. These images were captured using sonny digital high resolution color camera with 16 mega pixel resolution. The storage format of the images is jpeg format. The used dataset consisted of 2 categories of infected tomato leaves based on two disorders (100 images for each class) including powdery mildew and late blight that take hight ranked in its distorted effect on tomatoes crops between all tomato diseases.

The proposed SVM-based tomato diseases detection approach has been implemented considering the following steps: the the approach started with preprocessing phase that used to prepare input image to hight processing steps. It began with extracting every tomatoes leaf and isolated it into single image by cropping it using adobe photo shop 6. Since the input image was taken and saved in a very hight resolution, that may increase computation time, so the input image has been resized to 512 x 512 pixel and background of the image has been removed using background subtraction technique with some morphological operations. In this paper, gabor filters have been used to extract important features that perfectly describe the input images. Each of the filters in the filter bank is implemented as a 8 x 8 convolution mask for each of its real and imaginary components.

The acquired images were converted to HSV color space and 6 components of the image (R,G,B,H,S,V) have been extracted. To construct feature vector of each image components; a vector of 64 length was obtained from the average output for every $i$th filter. Vector of 3 length consisted of: cost function $J(i)$, maximum average output $D^i_{max}$ and minimum average output $D^i_{min}$. At the end of this step, feature vector of (64+3)x6= 402 length that describe the image has been obtained.

The dataset is of total 200 images were used for both training and testing datasets with 10-fold cross-validation. Training dataset is divided into 2 classes representing the different types of tomato diseases, 3/4 of this dataset was used in training phase, while the left was used in test phase. SVMs algorithm was employed with different kernel functions including Cauchy kernel, Invmult Kernel and Laplacian Kernel. In connection with the best set of hyperparameters selection for the different kernel functions, grid search technique has been used based on nested cross validation. Searching has been started for the optimal parameters in the big scale, then the searching space was narrowed down until satisfied.

Receiver operating characteristic, or in simply ROC curve, is a graphical plot that illustrates the performance of a binary classifier system. It is created by plotting the fraction of true positives out of the positives (sensitivity) versus the fraction of false positives out of the negatives (one specificity), at various threshold settings [18]. To obtain more accuracy in the evaluation step, ROC curve is plotted for SVM to illustrate its performance. Figure 2 shows receiver operating characteristic (ROC) curve and area under curve (AUC) for the best result from different kernel functions using SVM approach with 10-fold cross-validation and total of 150 images used for training and 50 images used for testing. ROC curve that represents best feature result for Invmult kernel function is represented in figure 3a, while figure 3b illustrated ROC curve for best result of Laplacian Kernel function. since classification accuracy result of Cauchy kernel function equal 100 %, so drowning its ROC curve is useless.

In order to visualize the high dimensional data, we apply MDS to the 402D data and reduce the dimension to 2D. Figure 4 illustrated classification result after dimention reduction.

## V. Conclusion and future works

In this paper, an effective SVM-based tomato diseases detection approach is proposed to recognize two types of tomatoes leaf diseases. The datasets of total 200 infected tomato leaf images with Powdery mildew and early blight were used for both training and testing phase. in this paper SVM is employed using different kernel functions including Cauchy kernel, Invmult Kernel and Laplacian Kernel. Grid search and N-fold cross-validation techniques were used to parameters selection and performance evaluation of the presented approach.

At first glance, the experimental results show that the overall accuracy offered by the suggested approach is very high using cauchy and Laplacian kernel functions 100% and 98% when comparing with Invmult kernel 78%.

## REFERENCES


[1] Peralta, E. Iris , and M. David Spooner. "History, origin and early cultivation of tomato (Solanaceae)." *Genetic improvement of Solanaceous crops*, vol. 2, pp. 1-27, 2006.

[2] Moshou, Dimitrios, C. Bravo, Jonathan West, Stijn Wahlen, Alastair McCartney, and Herman Ramon. "Automatic detection of yellow rustin wheat using reflectance measurements and neural networks." *Computers and electronics in agriculture*, vol. 44, no. 3, pp. 173-188, 2004.

[3] Zhang, Yan-Cheng, Han-Ping Mao, Bo Hu, and Ming-Xi Li. "Features selection of cotton disease leaves image based on fuzzy feature selection techniques." In: *proceeding of the ICWAPR'07 International Conference on Wavelet Analysis and Pattern Recognition*, vol. 1, pp. 124-129, 2007.

[4] P. Sanyal, and S. C. Patel. "Pattern recognition method to detect two diseases in rice plants." *The Imaging Science Journal*, vol. 56, no. 6, pp. 319-325, 2008.

[5] Huang, Kuo-Yi. "Application of artificial neural network for detecting Phalaenopsis seedling diseases using color and texture features." *Computers and Electronics in agriculture*, vol. 57, no. 1, pp. 3-11, 2007.

[6] S. G. Mallat. "A theory for multiresolution signal decomposition: the wavelet representation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674-693, 1989.

[7] Kong, Wai Kin, David Zhang, and Wenxin Li. "Palmprint feature extraction using 2-D Gabor filters." *Pattern recognition*, vol. 36, no. 10, pp. 2339-2347, 2003.

[8] Gabor, Dennis. "Theory of communication. Part 1: The analysis of information." *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429-441, 1946.

[9] K. Chui Charles. "An introduction to wavelets wavelet analysis and its applications." *Academic, Boston*, 1992.

[10] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning*, vol. 20, no. 3, pp. 273-297, 1995.

[11] Chen, Hui-Ling, Bo Yang, Gang Wang, Su-Jing Wang, Jie Liu, and Da-You Liu. "Support vector machine based diagnostic system for breast cancer using swarm intelligence." *Journal of medical systems*, vol. 36, no. 4, pp. 2505-2519, 2012.

[12] Aboul Ella Hassanien, Tai-hoon Kim. "Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks." *Journal of Applied Logic*, vol. 10, no. 4, pp. 277-284, 2012.

[13] G. Schaefer, Aboul Ella Hassanien, J. Jiang. "Computational Intelligence in Medical Imaging: Techniques and Applications." *CRC Press*, 2008.

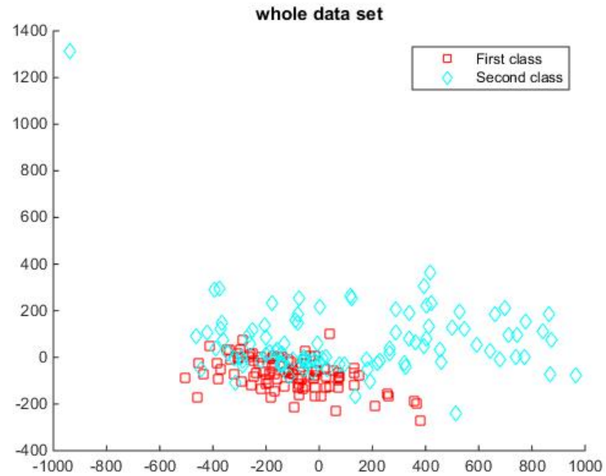
[14] Kumar, Ajay, and Grantham KH Pang. "Defect detection in textured materials using Gabor filters." *IEEE Transactions on Industry Applications*, vol. 38, no. 2, pp. 425-440, 2002.

[15] Manjunath, Bangalore S., and Wei-Ying Ma. "exture features for browsing and retrieval of image data." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837-842, 1996.

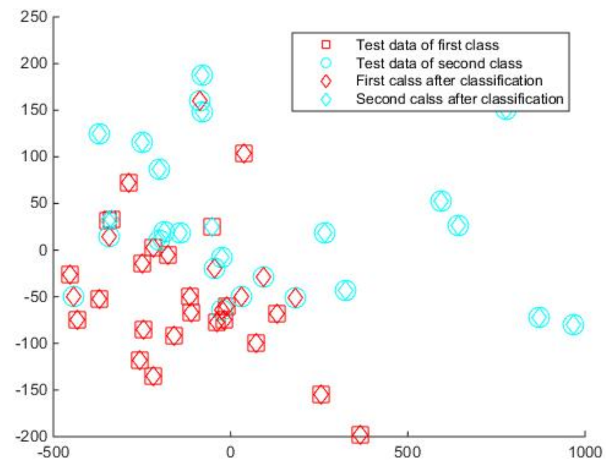
[16] Arlot, Sylvain, and Alain Celisse. "A survey of cross-validation procedures for model selection." *Statistics surveys*, vol. 4, pp. 40-79, 2010.

[17] Granholm, Viktor, William S. Noble, and Lukas Kll. "A cross-validation scheme for machine learning algorithms in shotgun proteomics." *BMC bioinformatics 13.Suppl*, vol. 16, 2012.

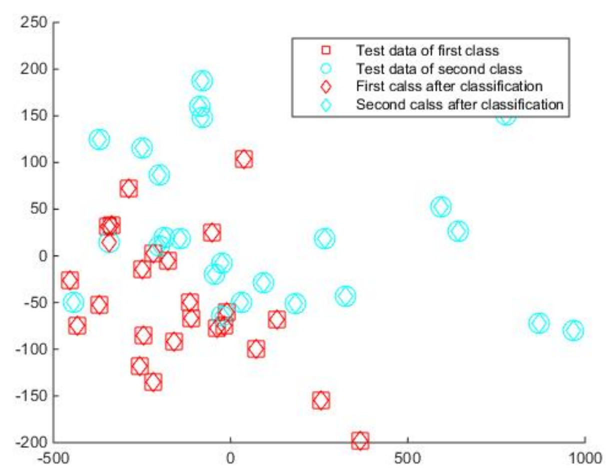
[18] ElBedwehy, Mona Nagy, Mohamed Elsayed Ghoneim, Aboul Ella Hassanien, and Ahmad Taher Azar. "A computational knowledge representation model for cognitive computers." *DD*, vol. 25, no. 7-8, pp. 1517-1534, 2014.


(a) whole dataset




(b) Classification result for invmult kernel SVM




(c) Classification result for Laplacian kernel SVM