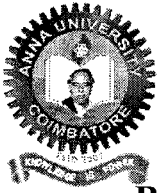


0 - 2551



BLOOD CELL RECOGNITION USING NEURAL NETWORKS

by

V.BIBIN CHRISTOPHER

Reg. No : 0720108004

of

KUMARAGURU COLLEGE OF TECHNOLOGY

COIMBATORE – 641 006

**(AN AUTONOMOUS INSTITUTION AFFILIATED TO ANNA UNIVERSITY
COIMBATORE)**

A PROJECT REPORT

Submitted to the

FACULTY OF INFORMATION AND COMMUNICATION ENGINEERING

**In partial fulfillment of the requirements
for the award of the degree**

Of

MASTER OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

MAY, 2009

BONAFIDE CERTIFICATE

Certified that this project report titled “**BLOOD CELL RECOGNITION USING NEURAL NETWORKS**” is the bonafide work of **Mr.V.BIBIN CHRISTOPHER (0720108004)** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other project report of dissertation on the basis of which a degree or ward was conferred on an earlier occasion on this or any other candidate.



GUIDE

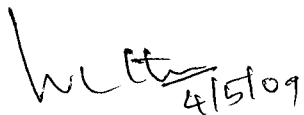
(Mrs. D.CHANDRAKALA)



HEAD OF THE DEPARTMENT

(Dr.S.THANGASAMY)

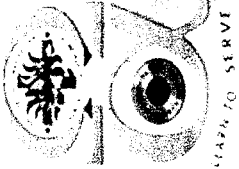
The candidate with **University Register No. 0720108004** was examined by us in Project Viva-Voce examination held on 04.05.09



INTERNAL EXAMINER



EXTERNAL EXAMINER



Departments of Computer Science and Engineering & Information Technology

CSI COLLEGE OF ENGINEERING



Ketti, The Nilgiris- 643 215

THIRD NATIONAL CONFERENCE ON EMERGING TECHNOLOGIES - 2009

CERTIFICATE

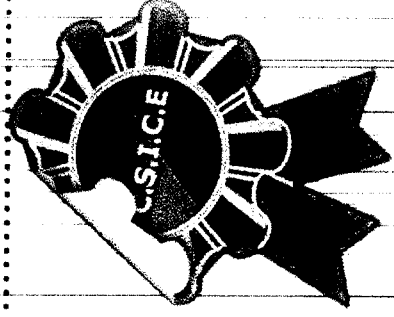
This is to certify that **Dr./Prof./Mr./Ms. BIBIN CHRISTOPHER . V**

KUMARAGURU COLLEGE OF TECHNOLOGY

has participated in
CET '09 and presented a paper entitled **BLOOD CELL RECOGNITION USING**
NEURAL NETWORKS

at the conference held on 11th March 2009.

Authors: **Mr. V. Bibin Christopher, Ms. Chandrakala . D**



CO-ORDINATOR

CONVENER

PRINCIPAL

ABSTRACT

The recognition of the blast cells in the bone marrow of the patients suffering from leukemia disease is characterized by severe illness and the abnormal growth of leukocytes. The recognition of this disease at the development stage of the illness and proper treatment should be provided to the patients. There are different cell lines in the bone marrow namely the megacaryocytic, erythrocytic, monocytic, lymphocytic and granulocytic. To the most known and recognized cells belong: monoblasts, promonocytes, monocytes, myeloblasts, promyelocytes, myelocyte, metamyelocytes, proerythroblasts, basophilic erythroblast, polychromatic erythroblast, pyknotic erythroblast

Up to now no automatic system exists that could recognize the blood cells with the accuracy comparable to the human expert. Although some attempts have been presented to solve this problem and the results are still not satisfactory when compared to the efficiency of the human expert. In this project an automatic blood cell recognition system is designed and it uses support vector machine as classifier.

Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. By using support vector machine the classification is done and the error rate is calculated for the image taken from the patients. From the error rate one can easily identify how much percentage the patients are affected by leukemia disease and the treatment can be done accordingly

ஆய்வு சுருக்கம்

இந்த ஆய்வில் எலும்பு மஜ்ஜையில் பல வகையான செல்கள் உள்ளன. அவை மெகாகேரியாடிக், எரித்ரோசைட்டிக், மோனோசைட்டிக், லைம்போசைட்டிக் மற்றும் கிரானிலோசைட்டிக். மிக நுட்பமான இரத்த செல்களைக் கண்டறிவதில் மனித அறிவுக்கு இணையான எந்த ஒரு தானியங்கி அமைப்பும் (இயந்திரமும்) முடிவுகளை தருவதில்லை.

இந்த திட்ட அறிக்கையில் தானியங்கி இரத்த செல் அமைப்பு வடிவமைக்கப்பட்டு சப்போட் வெக்டர் இயந்திரம் பயன்படுத்தப்படுகிறது.

சப்போட் வெக்டர் இயந்திரம் ஓர் வகைப்படுத்தும் கருவி. சப்போட் வெக்டர் இயந்திரத்தை பயன்படுத்தி நோயாளிகளின் பிம்பத்தில் உள்ள பிழை அளவுகளையும் மற்றும் அளவற்ற வகைப்படுத்தலாம் . அந்த பிழை அளவுகளிலிருந்து நோயாளி அந்த அளவிற்கு பாதிக்கப்பட்டிருக்கிறார் என்பதை சுலபமான முறையில் கண்டறிந்து அதற்கு தகுந்தார்போல் சிகிச்சை அளிக்கலாம்.

ACKNOWLEDGEMENT

I express my profound gratitude to our Chairman **Padmabhusan Arutselver Dr. N. Mahalingam B.Sc, F.I.E** for giving this great opportunity to pursue this course.

I would like to begin by thanking to **Dr. Joseph V. Thanikal, Ph.D.,** *Principal* and Vice Principal **Prof. R. Annamalai,** *Principal*, for providing the necessary facilities to complete my thesis.

I take this opportunity to thank **Dr.S.Thangasamy, Ph.D.,** *Head of the Department,* Computer Science and Engineering, for his precious suggestions.

I register my hearty appreciation to **Mrs. D.Chandrakala M.E.,** *my thesis advisor.* I thank for her support, encouragement and ideas. I thank her for the countless hours she has spent with me, discussing everything from research to academic choices.

I thank all project committee members for their comments and advice during the reviews. Special thanks to **Mrs.V.Vanitha M.E.,** *Assistant professor,* Department of Computer science and Engineering, for arranging the brain storming project review sessions.

I would like to convey my honest thanks to all **Teaching** staff members and **Non Teaching** staffs of the department for their support. I would like to thank all my classmates who gave me a proper light moments and study breaks apart from extending some technical support whenever I needed them most.

I dedicate this project work to my **parents** for no reasons but feeling from bottom of my heart, without their love this work wouldn't possible.

TABLE OF CONTENTS

Contents	Page No.
Abstract	iii
Abstract (Tamil)	iv
List of Figures	viii
List of Tables	x
1. Introduction	1
2. Details of Literature Survey	
2.1 Feature Generation for the cell image Recognition of Myelogenous Leukemia	6
2.2 Recognition of blood and bone marrow cells	7
2.3 System-Level Training of Neural Networks for counting White blood cells	8
2.4 An Introduction to Variable and Feature Selection	9
2.5 A Medical Image Segmentation method Based on Watershed Transform	11
2.6 Support Vector Machines applied to White Blood Cell	11
2.7 Feature Extraction And Classification Of Blood Cells	12
2.8 Blood Cell Identification Using Neural Networks	13
3. Design of Automatic Blood cell Recognition System	
3.1 Digital Camera	15
3.2 Feature Generation	16
3.3 Cell Extraction	17
3.3.1 Structuring Elements	18
3.3.2 Eroding an Image	20

3.3.3 Dilating an Image	21
3.3.4 Opening	23
3.3.5 Closing	24
4. Feature Selection Techniques	
4.1 Selection based on Mean and Variance	25
4.1.1 Standard Deviation	25
4.1.2 Variance	26
4.1.3 Coefficient of Variation	27
4.2 Correlation Analysis	28
5. Support Vector Machine	
5.1 Introduction	30
5.2 Statistical Learning Theory	31
5.3 SVM for Classification	35
5.4 Properties of SVM	36
5.5 SVM Applications	36
5.6 Weakness of SVM	37
6. Experimental Results and Discussion	38
7. Conclusion and Future Work	47
8. Appendices	
8.1 Appendix I	48
8.2 Appendix II	52
9. References	53

LIST OF FIGURES

FIGURE	TITLE	PAGE NO
Fig 3.1	Proposed Automatic Blood Cell Recognition system	16
Fig.3.2	Bone Marrow Image	19
Fig 3.3	Origin of a Diamond-Shaped Structuring Element	20
Fig 3.4	Erosion of Image	22
Fig 4.1	Perfect positive correlation	29
Fig 4.2	Perfect negative correlation	29
Fig 5.1	Simple Neural Network Multilayer Perceptron	33
Fig 5.2	Hyperplanes to classify Data	33
Fig 5.3	illustration of Linear SVM	34
Fig 5.4	Representation of Hyper planes	35
Fig 5.5	Representation of Support Vectors	36
Fig 6.1	Original RGB Image fed for Segmentation	39
Fig 6.2	Converted Gray Scale Image	40
Fig 6.3	Cells obtained after Morphological Operations	40
Fig 6.4	Extracted Image	41
Fig 6.5	The Extracted Features of individual Cell	42
Fig 6.6	Non- Ranked Features and calculation of Mean and Variance	42
Fig 6.7	Ranked Features for Features Selection	43
Fig 6.8	Performance characteristics of cell between 100 & 200	44

Fig 6.9	Performance characteristics of cell between 100 & 250	45
Fig 6.9	Performance characteristics of cell between 100 & 400	46
Fig 6.11	Performance characteristics of cell between 100 & 250	47

LIST OF TABLES

TABLE	TITLE	PAGE NO
Table 6.1	Representation of cells between 100 & 200	44
Table 6.2	Representation of cells between 100 & 250	45
Table 6.3	Representation of cells between 100 & 400	46
Table 6.4	Representation of cells between 10 & 400	47

CHAPTER 1

INTRODUCTION

A disease characterized by a progressive and abnormal accumulation of white blood cells, or leukocytes. Leukemic cells are malignant because they have three characteristics common to all cancers:

1. They exhibit uncontrolled growth that is frequently associated with an inability to mature normally
2. They arise from a single precursor cell
3. They disregard anatomic boundaries and metastasize to organs or tissues where leukocytes are not normally found.

Normal leukocytes are grouped into two primary types or lineages, myeloid and lymphoid, and virtually any cell of either lineage can become leukemic. Leukemias are also divided into broad categories that are based on the cell involved (myeloid or lymphoid) and disease aggressiveness (either acute or chronic).

The two major types of leukemia usually differ in signs and symptoms.

1. Acute leukemia

It has a relatively rapid onset, and those with the disease often experience problems immediately.

2. Chronic leukemia

It has an insidious course and is frequently discovered during an examination for an unrelated problem.

ACUTE LEUKEMIA

In acute leukemia, the maturation process of the white blood cells is interrupted. The immature cells (or "blasts") proliferate rapidly and begin to accumulate in various organs and tissues, thereby affecting their normal function.

This uncontrolled proliferation of the immature cells in the bone marrow affects the production of the normal red blood cells and platelets as well.

Acute leukemias are of two types:

1. Acute lymphocytic leukemia

In acute lymphocytic leukemia (ALL), the T or B lymphocytes become cancerous. The B cell leukemias are more common than T cell leukemias.

2. Acute myelogenous leukemia

It is also known as acute nonlymphocytic leukemia (ANLL), is a cancer of the monocytes and/or granulocytes.

Leukemias account for 2% of all cancers. Because leukemia is the most common form of childhood cancer, it is often regarded as a disease of childhood. According to the estimates of the American Cancer Society (ACS), approximately 29,000 new cases of leukemia were diagnosed in 1998. Internationally, leukemia is the fourth most common cancer among people age 15 to 19 years old.

CHRONIC LEUKEMIA

In chronic leukemias, the cancer starts in the blood cells made in the bone marrow. The cells mature and only a few remain as immature cells. However, even though the cells mature and appear normal, they do not function as normal cells.

Depending on the type of white blood cell that is involved, chronic leukemia can be classified as

1. Chronic lymphocytic leukemia

Chronic lymphocytic leukemia (CLL) involves the T or B lymphocytes. B cell abnormalities are more common than T cell abnormalities. T cells are affected in only 5% of the patients. The T and B lymphocytes can be differentiated from the other types of white blood cells based on their size and by

the absence of granules inside them. Chronic lymphocytic leukemia (CLL) often shows no early symptoms and may remain undetected for a long time.

2. Chronic myelogenous leukemia

In chronic myelogenous leukemia (CML), the cells that are affected are the granulocytes. Chronic myelogenous leukemia (CML), on the other hand, may progress to a more acute form.

CAUSES & SYMPTOMS

Leukemia strikes both sexes and all ages and its cause is mostly unknown. However, chronic leukemia has been linked to genetic abnormalities and environmental factors. For example, exposure to ionizing radiation and to certain organic chemicals, such as benzene, is believed to increase the risk for getting leukemia. Chronic leukemia occurs in some people who are infected with two human retroviruses (HTLV-I and HTLV-II). An abnormal chromosome known as the Philadelphia chromosome is seen in 90% of those with CML. The incidence of chronic leukemia is slightly higher among men than women.

The incidence of acute leukemia is slightly higher among men than women. A higher incidence of leukemia has also been observed among persons with Down syndrome and some other genetic abnormalities.

A history of diseases that damage the bone marrow, such as aplastic anemia, or a history of cancers of the lymphatic system puts people at a high risk for developing acute leukemias. Similarly, the use of anticancer medications, immunosuppressants, and the antibiotic chloramphenicol also are considered risk factors for developing acute leukemias.

The symptoms of leukemia are generally vague and non-specific. A patient may experience all or some of the following symptoms:

1. weakness or chronic fatigue
2. fever of unknown origin
3. weight loss that is not due to dieting or exercise
4. frequent bacterial or viral infections
5. headaches
6. skin rash
7. non-specific bone pain
8. easy bruising
9. bleeding from gums or nose
10. blood in urine or stools
11. enlarged lymph nodes and/or spleen
12. abdominal fullness

DIAGNOSIS

Like all cancers, leukemias are best treated when found early. There are no screening tests available. If there is no reason to suspect leukemia, he or she will conduct a thorough physical examination to look for enlarged lymph nodes in the neck, underarm, and pelvic region. Swollen gums, enlarged liver or spleen, bruises, or pinpoint red rashes all over the body are some of the signs of leukemia. Urine and blood tests may be ordered to check for microscopic amounts of blood in the urine and to obtain a complete differential blood count. This count will give the numbers and percentages of the different cells found in the blood. An abnormal blood test might suggest leukemia, however, the diagnosis has to be confirmed by more specific tests.

We may perform a bone marrow biopsy to confirm the diagnosis of leukemia. During the biopsy, a cylindrical piece of bone and marrow is removed, generally from the hip bone. These samples are sent to the laboratory for examination. In addition to

diagnosis, the biopsy is also repeated during the treatment phase of the disease to see if the leukemia is responding to therapy.

A spinal tap (lumbar puncture) is another procedure to diagnose leukemia. In this procedure, a small needle is inserted into the spinal cavity in the lower back to withdraw some cerebrospinal fluid and to look for leukemic cells.

Standard imaging tests, such as x rays, computed tomography scans (CT scans), and magnetic resonance imaging (MRI) may be used to check whether the leukemic cells have invaded other areas of the body, such as the bones, chest, kidneys, abdomen, or brain. A gallium scan or bone scan is a test in which a radioactive chemical is injected into the body. This chemical accumulates in the areas of cancer or infection, allowing them to be viewed with a special camera.

CHAPTER 2

DETAILS OF LITERATURE SURVEY

The preprocessing methods of the leukemic blast cells image in order to generate the features well characterizing different types of cells. The segmentation of the bone marrow image in to individual cells by the morphological operations. These features are used as the input signals applied to the support vector machine used as the classifier. The details of literature review are as follows

2.1 FEATURE GENERATION FOR THE CELL IMAGE RECOGNITION OF MYELOGENOUS LEUKEMIA

The preprocessing methods of the leukemic blast cells image in order to generate the features well characterizing different types of cells. The solved problems include: the segmentation of the bone marrow aspirate by applying the watershed transformation, selection of individual cells, feature generation on the basis of texture, statistical and geometrical analysis of the cells. These features are used as the input signals applied to the support vector machine used as the classifier. The numerical results of recognition of 10 different cell types are presented.

The acute leukemia is a disease of the leukocytes and their precursors. It is characterized by the appearance of immature, abnormal cells in the bone marrow and peripheral blood. The aspirated marrow is found to be infiltrated by abnormal cells.

There are different cell types in the bone marrow. The most known and recognized abnormal cells include monoblasts, promonocytes, monocytes, myeloblasts, promyelocytes, myelocytes, metamyelocytes, proerythroblasts, basophilic erythroblasts, polychromatic erythroblasts, orthochromatic erythroblasts, lymphocytes, plasmocytes, megacaryoblasts, megacaryocytes, etc. The variety of cells occurring in the bone marrow demands a high expertise of the analyst, which is usually verbal one. For improving the reliability of the analysis and diagnosis, computer based digital image processing offers a useful tool.

2.2 RECOGNITION OF BLOOD AND BONE MARROW CELLS

This paper presents a novel cell classification method based on image retrieval by learning with kernel. Cell image is firstly segmented into cytoplasm and nucleus regions in order to keep more spatial information. RGB color histogram of cell and two intensity histograms corresponding to those local regions compose feature vector represents the cell image. Kernel principal component analysis (KPCA) is utilized to extract effective features from the feature vector. The weight coefficients of features are estimated automatically using relevance feedback strategy by linear support vector machine (SVM). Classification depends on the decision distance obtained by SVM and the nearest center criterion. Experimental results on the ten-class task of 400 cells from blood and bone marrow smears show a 90.5% classification accuracy of the method when combined with standardized sample preparation and image acquisition.

The analysis of blood and bone marrow smears is a powerful diagnostic tool for the detection of leukemia. It is a classical and challenging pattern recognition task that always includes two stages: one is object detection/segmentation and the other is object recognition/classification. In this paper, we focus on the solution of the later. There are many types of cell with different lineage and maturity level in bone marrow. Most of them only have subtle visible differences. It is difficult to achieve a consistent diagnose during microscopic evaluation by subjective impressions of observers. Computer-assisted morphologic cell classification can improve accuracy, objectivity and reproducibility in diagnosis. Yet, the classification performance extremely lies on the strategy of image feature selection and extraction.

A CBIR-based method to classify complex cells from blood and bone marrow smears. Color and intensity histograms as image-based feature representation are mainly adopted in our method, and kernel principal component analysis (KPCA) is used to reduce the high dimensionality of the data representation and deal with the

nonlinear distribution of features. In order to achieve satisfied retrieval accuracy and generalization performance, a relevance feedback strategy based on linear SVM is presented for training of classifiers. We analyze 10-class task involving monocytic and granulocytic series of white blood cells with new method. Experimental results show that over 90% accuracy could be achieved.

2.3 SYSTEM-LEVEL TRAINING OF NEURAL NETWORKS FOR COUNTING WHITE BLOOD CELLS

Neural networks (NNs) that are trained to perform classification may not perform as well when used as a module in a larger system. In this correspondence, we introduce a novel, system-level method for training NNs with application to counting white blood cells. The idea is to phrase the objective function in terms of total count error rather than the traditional class-coding approach because the goal of this particular recognition system is to accurately count white blood cells of each class, not to classify them. An objective function that represents the sum of the squared counting errors (SSCE) is defined. A batch-mode training scheme based on back-propagation and gradient descent is derived. Sigma and crisp counts are used to evaluate the counting performance. The testing results show that the network trained to minimize SSCE performs better in counting than a classification network with the same structure even though both are trained a comparable number of iterations.

Relative counts of different classes of white blood cells in bone marrow aid in the diagnosis of diseases, such as leukemia. According to the myelocytic or granulocytic series, white blood cells in human bone marrow are classified into six discrete classes in accordance with their ages, namely, Myeloblast, Promyelocyte, Myelocyte, Metamyelocyte, Band, and PMN, respectively, ordered from the youngest to the oldest. When a white blood cell becomes older, several features change. For example, its size is smaller, its nucleus shape changes from round-shaped to segments, and its texture is coarser. Sample images of all six cell classes are shown in Fig. 1. In an effort to overcome the tedious and time-consuming task of human

experts in counting white blood cells in bone marrow or peripheral blood, many automated techniques have been proposed. Although some commercial products have been introduced for peripheral blood, the procedure has not been automated for bone marrow due to the complexity of the images.

2.4 AN INTRODUCTION TO VARIABLE AND FEATURE SELECTION

Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. These areas include text processing of internet documents, gene expression array analysis, and combinatorial chemistry. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. The contributions of this special issue cover a wide range of aspects of such problems: providing a better definition of the objective function, feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods.

One is gene selection from microarray data and the other is text categorization. In the gene selection problem, the variables are gene expression coefficients corresponding to the abundance of mRNA in a sample (e.g. tissue biopsy), for a number of patients. A typical classification task is to separate healthy patients from cancer patients, based on their gene expression profile. Usually fewer than 100 examples (patients) are available altogether for training and testing. But, the number of variables in the raw data ranges from 6000 to 60,000. Some initial filtering usually brings the number of variables to a few thousand. Because the abundance of mRNA varies by several orders of magnitude depending on the gene, the variables are usually standardized. In the text classification problem, the documents are represented by a "bag-of-words", that is a vector of dimension the size of the vocabulary containing word frequency counts (proper normalization of the variables also apply).

Vocabularies of hundreds of thousands of words are common, but an initial pruning of the most and least frequent words may reduce the effective number of words to 15,000. Large document collections of 5000 to 800,000 documents are available for research. Typical tasks include the automatic sorting of URLs into a web directory and the detection of unsolicited email (spam).

There are many potential benefits of variable and feature selection: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance. Some methods put more emphasis on one aspect than another, and this is another point of distinction between this special issue and previous work. The papers in this issue focus mainly on constructing and selecting subsets of features that are useful to build a good predictor. This contrasts with the problem of finding or ranking all potentially relevant variables. Selecting the most relevant variables is usually suboptimal for building a predictor, particularly if the variables are redundant. Conversely, a subset of useful variables may exclude many redundant, but relevant, variables. This introduction surveys the papers presented in this special issue.

The depth of treatment of various subjects rejects the proportion of papers covering them: the problem of supervised learning is treated more extensively than that of unsupervised learning; classification problems serve more often as illustration than regression problems, and only vectorial input data is considered. Complexity is progressively introduced throughout the sections: The first section starts by describing filters that select variables by ranking them with correlation coefficients (Section 2). Limitations of such approaches are illustrated by a set of constructed examples (Section 3). Subset selection methods are then introduced (Section 4). These include wrapper methods that assess subsets of variables according to their usefulness to a given predictor. We show how some embedded methods implement the same idea, but proceed more efficiently by directly optimizing a two-part objective function with a goodness-of-fit term and a penalty for a large number of

variables. We then turn to the problem of feature construction, whose goals include increasing the predictor performance and building more compact feature subsets

2.5 A MEDICAL IMAGE SEGMENTATION METHOD BASED ON WATERSHED TRANSFORM

The watershed transform is a popular segmentation method coming from the field of mathematical morphology. In order to solve the over-segmentation of watershed algorithm's applied to medical images, we first transform the original image into a morphological gradient image by using morphology methods and treat it with an opening-closing filter bank to get a reference image with less noise interference; then we segment the reference image by using the watershed algorithm; finally, according to the test criterion with respect to the gray mean of the region boundary and the region area, we merge the segment field. Experiments applying to segmentation of medical images by using watershed algorithm show that the image processing before and after segmentation prevents the over-segmentation effectively.

Image segmentation is an important step toward the analysis phase in many medical image-processing tasks such as image guidance operation, tumors radiotherapy, evaluation of therapies and so on, which are supposed to have an exact segmentation on image. The results of medical image segmentation are crucial to give a correct diagnosis.

2.6 SUPPORT VECTOR MACHINE FOR EFFICIENT BLOOD CELL RECOGNITION

The paper presents application of the genetic algorithm and Support Vector Machine for the recognition of the blood cells on the basis of image of the bone marrow aspirate. The automatic recognition system has been developed and the results of its numerical verification are presented and discussed. They show that

application of the genetic algorithm is a powerful tool for selection of the diagnostic features and leads to the significant increase of the accuracy of the whole system.

The relative counting and assessment of the blood cells of the bone marrow of the patients are very informative in clinical practice. It is especially important for patients suffering from leukemia in the observation of the development stage of the illness and at the diagnosing the treatment of the patients. The white blood cells in the human bone marrow are classified according to their age. We can recognize the myeloblast, promyelocyte, myelocyte, metamyelocyte, neutrophilic band and segmented. To put proper diagnosis of the disease we have to recognize the cells at different stages of their development and calculate their quantity in the aspirated bone marrow.

2.7 FEATURE EXTRACTION AND CLASSIFICATION OF BLOOD CELLS FOR AN AUTOMATED DIFFERENTIAL BLOOD COUNT SYSTEM

The differential blood counter (DBC) system that we have developed is an attempt to automate the task performed manually by experts in routine. Feature extraction and classification are two important components of our automated system. In this paper, classification of blood cells using various approaches including neural network based classifiers and support vector machine are presented together with the features used in the classification.

White cell composition of the blood reveals important diagnosis information about the patients as well as patient follow-up. The hematologist requires two types of blood count for diagnosis and screening. The first one is called the Complete Blood Count (CBC) and the second one is called the Differential Blood Count (DBC). CBC could be done by instruments called cytometer and could successfully be performed automatically. On the other hand, DBC is more reliable but currently it is a manual procedure to be done by hematology experts using microscope. In DBC, an expert counts 100 white blood cells on the smear at hand and computes the percentage of occurrence of each type of cell counted. The results reveal important information

about patient's health status. Apparently, DBC is a time consuming task that requires expert examination. Our automated differential blood counter system is an attempt for performing DBC automatically by the aid of statistical and neural network based classification methods.

The process of counting blood cells on smear images requires four steps. These steps are acquisition, segmentation, feature extraction, and classification. used in . Due to the fuzzy nature of the decision process in counting blood cells, a dedicated neural network counter is constructed in. In this work, the authors state the fact that the results of a counting session could be different between trials about 15%.

In order to conduct an automated counter, methods performing well for segmentation, feature extraction, and classification are needed. In our current system, segmentation is done by morphological preprocessing followed by the snake-balloon algorithm. Several types of features such as intensity and color based features, texture based features, and shape based features are utilized for a robust representation of the objects.

2.8 BLOOD CELL IDENTIFICATION USING NEURAL NETWORKS

The method of identifying three major blood cell types namely erythrocytes, leukocytes and platelets and to classify them based upon their morphological features using neural networks. The data are collected using peripheral blood smears from clinical patients. The image acquisition requires 100X magnification on all the blood smears, the preprocessing involves the use of median and edge enhance filters and the feature extraction is done by performing the wavelet transform on the images. Finally classification of the blood cell types is done using ALOPEX. Back Propagation trained neural networks. The efficacy of both networks is then compared by comparing their outputs and number of iterations required to reach the final result

There exists three major cellular constituents of the blood: **Erythrocytes or red cells** are non-nucleated biconcave diskettes with a diameter of about 8 μm , a thickness of about 2 μm at its edges, and a volume of about 94 μm^3 . The red cells

make up about 48 *YO* of the blood volume, approximately 5.2 million/pl of blood. **Leukocytes or white blood cells** are nucleated cells with diameters ranging from 6 to 20 μm . Normal blood contains between 4000 to 10,000 leukocytes/pl of blood. **Platelets** are cytoplasmic fragments of large cells called megakarocytes. They have a diameter of about 2-4 μm and normal blood contains between 150,000 to 350,000 platelets/pl of blood.

The most important function of the platelets is thrombosis and control of bleeding. The three cell types can be differentiated from each other based upon the following criterion: The size of all three cells is different i.e. they have different radii. Morphological features like the presence or absence of a nucleus in the cells and the shape of the nucleus and finally that the volume of the three cell **types** is different from one another.

The four major functions of peripheral blood analysis are to provide information for diagnosis to physicians, to provide data for selection of further pertinent tests to establish a diagnosis, to act as a guide to therapy and to act as an indicator of harmful effects of chemotherapy and radiotherapy. The wavelet transform, also known as the adaptive window analysis, was used in this project because it has excellent characteristic of localization both in time and frequency. Artificial neural networks (ANN) have been previously used in image classification. For this project the artificial neural networks were used because of their potential to serve as a real-time testing system.

CHAPTER 3

DESIGN OF AUTOMATIC BLOOD CELL RECOGNITION SYSTEM

The process of automatic recognition requires the extraction of individual blood cells, generation of diagnostic features and finally the recognition using chosen classifier. The applied automatic blood cell recognizing system is presented in the diagram.

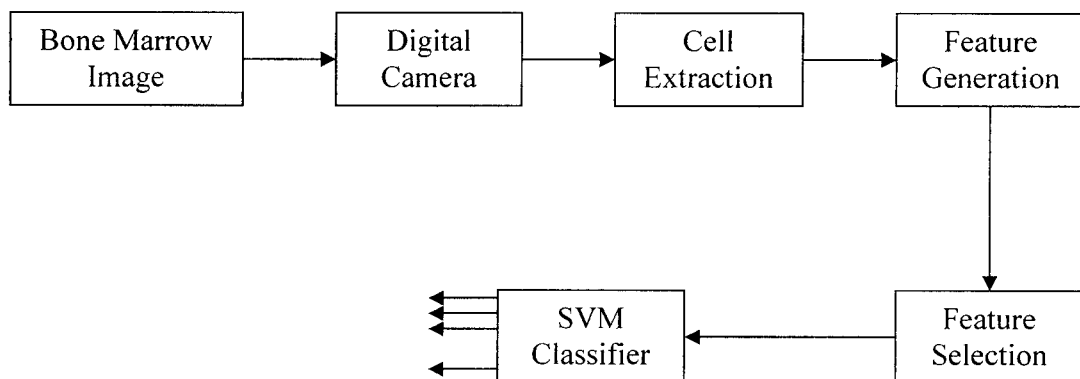


Fig 3.1: The proposed automatic blood cell recognition system

The bone marrow image is digitized using digital camera. The next step is the extraction of individual cells from the image. For each cell the feature generation and selection are performed. The selected features form the vector x applied to the input of Support Vector Machine (SVM) network working as the recognizing and classifying system.

3.1 DIGITAL CAMERA

The bone marrow smear samples have been collected from the patients suffering from leukemia. The acquired image was digitized using Olympus microscope with the magnification of 1000x and digital camera of resolution 1712x1368 pixels and the picture saved in RGB format. Individual cells were

detected using cell image segmentation. Ten features are extracted from each single-cell image. These features were extracted from each single-cell image without any preprocessing. Each white blood cell image was segmented to form a single-cell image

3.2 FEATURE GENERATION

The Main important feature of blood cell processing is the extraction of individual cells from the image of the bone marrow. This can be done by segmentation using morphological operations of watershed transformation.

Recognition of the blood cell on the basis of its image needs generation of the numerical features well describing the differences of images belonging to different classes. In characterizing the images by the numerical values we try to get the features strictly corresponding to these on the basis of which the human expert makes his diagnosis, that is the geometry of cell, texture, color and intensity of the image associated with different cell types.

Four families of features have been created.

The geometrical features include such parameters as radius, perimeter, area, the area of convex part of the cell, compactness, concavity, symmetry, major and minor axis lengths, etc. These parameters are determined only for the nucleus of the cell. Up to 19 geometrical features have been generated for each cell on the basis of these parameters.

The texture refers to an arrangement of the basic constituents of the material and in the digital image is depicted by the interrelationships between spatial arrangements of the image pixels. After some preliminary experiments, we have chosen two texture preprocessing methods, due to Unser and Markov random field. Up to 105 texture features have been generated for the cell image at normal and reduced resolutions.

The next set of features has been generated from the analysis of the intensity distribution of the image. The histograms of the image and gradient matrix of such intensity have been determined for R, G, B components of the image. On the basis of such analysis we have generated the following features: the mean and variance of the histogram of the image of nucleus and cytoplasm (separately) as well as for the gradient matrix of the image, the skewness and kurtosis of the image of the whole cell as well as for the gradient matrix of the whole cell. Up to 24 statistical features have been generated in this way for two colours (red and green).

The last set of features is related to the morphological operations performed on the images (erosion, dilation, opening and closing). These parameters include the area and number of separated objects of the image after application of some morphological operations. Up to 16 morphological parameters have been generated in this way. All features have been normalized, dividing their original values by the corresponding maxima.

3.3 CELL EXTRACTION

Image segmentation is a division of the image into different regions, each having certain properties. In a segmented image, the picture elements are no longer the pixels, but connected set of pixels, all belonging to the same region. We will use the Segmentation techniques to separate the individual cells from the set of cells creating the image. The recognition and separation of individual cells from the image of the blood cells is a very difficult task, since different regions are of little grey level variations and the borders of individual cells are hardly visible.

The individual cells are close to each other and the borders among them are not well defined. The task of segmentation of the image is focused on the automatic recognition and separation of each cell for further processing, in order to obtain stable features, useful in recognition of the cell. In solving the segmentation task we have used the morphological operations.

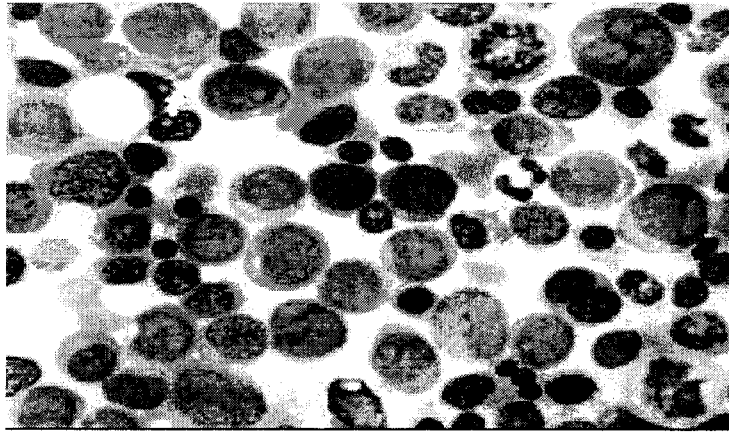


Fig.3.2: Bone Marrow Image

The morphological operations aim at extracting relevant structures of the image by probing the image with another set of a known shape called structuring element, chosen as the result of prior knowledge concerning the geometry of the relevant and irrelevant image structures.

The most known morphological operations are

1. Erosion
2. Dilation
3. Opening
4. Closing

3.3.1 Structuring Elements

An essential part of the dilation and erosion operations is the structuring element used to probe the input image. Two-dimensional, or *flat*, structuring elements consist of a matrix of 0's and 1's, typically much smaller than the image being processed. The center pixel of the structuring element, called the *origin*, identifies the pixel of interest--the pixel being processed. The pixels in the structuring element containing 1's define the *neighborhood* of the structuring element. These pixels are also considered in the dilation or erosion processing. Three dimensional, or *nonflat*, structuring elements use 0's and 1's to define the extent of the structuring element in the *x*- and *y*-plane and add height values to define the third dimension.

The Origin of a Structuring Element

The morphological functions use this code to get the coordinates of the origin of structuring elements of any size and dimension.

```
origin = floor((size(nhood)+1)/2)
```

(In this code, `nhood` is the neighborhood defining the structuring element. Because structuring elements are MATLAB objects, you cannot use the size of the `STREL` object itself in this calculation. You must use the `STREL.getnhood` method to retrieve the neighborhood of the structuring element from the `STREL` object. For information about other `STREL` object methods, see the [strel](#) function reference page.)

For example, the following illustrates a diamond-shaped structuring element.

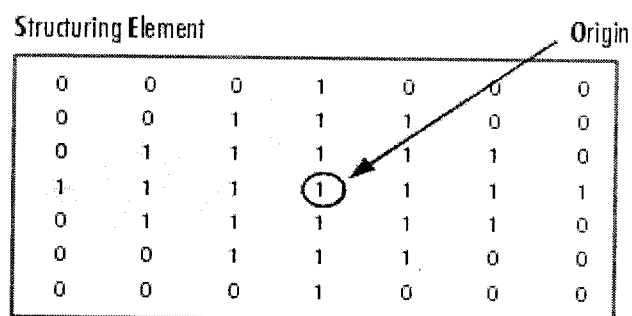


Fig. 3.3: Origin of a Diamond-Shaped Structuring Element

3.3.2 Eroding an Image

To erode an image, use the `imerode` function. The `imerode` function accepts two primary arguments:

- The input image to be processed (grayscale, binary, or packed binary image)

- A structuring element object, returned by the `strel` function, or a binary matrix defining the neighborhood of a structuring element

`imerode` also accepts three optional arguments: `PADOPT`, `PACKOPT`, and `M`.

The `PADOPT` argument affects the size of the output image. The `PACKOPT` argument identifies the input image as packed binary. If the image is packed binary, `M` identifies the number of rows in the original image. (See the [bwpack](#) reference page for more information about binary image packing.)

The following example erodes the binary image, `circbw.tif`:

1. Read the image into the MATLAB workspace.

```
BW1 = imread('circbw.tif');
```

2. Create a structuring element. The following code creates a diagonal structuring element object. (For more information about using the `strel` function, see [Structuring Elements](#).)

```
SE = strel('arbitrary',eye(5));
SE=
```

```
Flat STREL object containing 5 neighbors.
```

```
Neighborhood:
```

```

1     0     0     0     0
0     1     0     0     0
0     0     1     0     0
0     0     0     1     0
0     0     0     0     1
```

3. Call the `imerode` function, passing the image, `BW`, and the structuring element, `SE`, as arguments.

```
BW2 = imerode(BW1,SE);
```

1. Notice the diagonal streaks on the right side of the output image. These are due to the shape of the structuring element.

```
imshow(BW1)
figure, imshow(BW2)
```

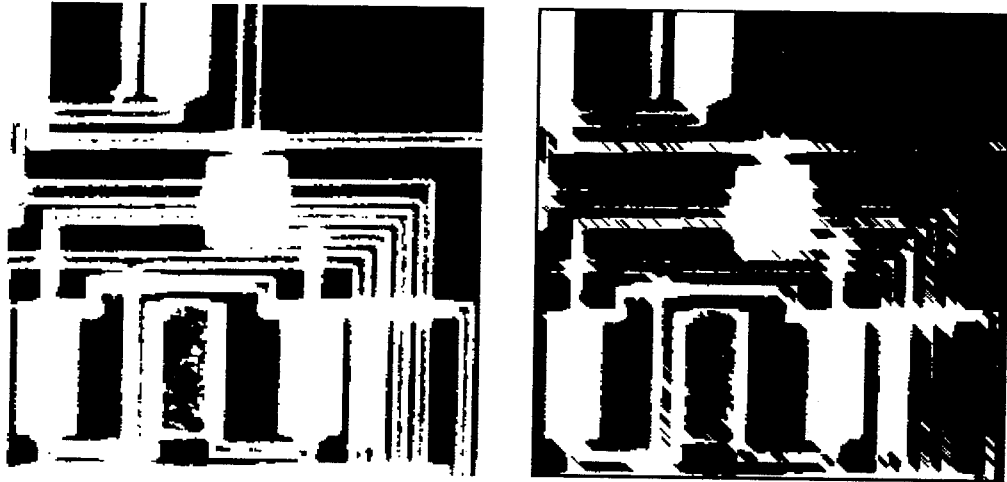


Fig.3.4: Erosion of image

The basic effects:

1. Erode away the boundary regions of foreground pixels.
2. Areas of foreground pixels shrink in size
3. holes within those regions become larger

3.3.3 Dilating an Image

To dilate an image, use the `imdilate` function. The `imdilate` function accepts two primary arguments:

1. The input image to be processed (grayscale, binary, or packed binary image)
2. A structuring element object, returned by the `strel` function, or a binary matrix defining the neighborhood of a structuring element

`imdilate` also accepts two optional arguments: `PADOPT` and `PACKOPT`. The `PADOPT` argument affects the size of the output image. The `PACKOPT` argument identifies the input image as packed binary. (See the [bwpack](#) reference page for information about binary image packing.)

This example dilates a simple binary image containing one rectangular object.

```
BW = zeros(9,10);
BW(4:6,4:7) = 1
BW =
    0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0
    0     0     0     1     1     1     1     0     0     0
    0     0     0     1     1     1     1     0     0     0
    0     0     0     1     1     1     1     0     0     0
    0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0
```

To expand all sides of the foreground component, the example uses a 3-by-3 square structuring element object. (For more information about using the `strel` function, see [Structuring Elements](#).)

```
SE = strel('square',3)
SE =

Flat STREL object containing 3 neighbors.

Neighborhood:
    1     1     1
    1     1     1
    1     1     1
```

To dilate the image, pass the image, `BW`, and the structuring element, `SE`, to the `imdilate` function. Note how dilation adds a rank of 1's to all sides of the foreground object.

```
BW2 = imdilate(BW,SE)
```

BW2 =

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

The basic effects:

1. Gradually enlarge the boundaries of regions of foreground pixels (i.e. white pixels, typically).
2. Areas of foreground pixels grow
3. holes within those regions become smaller

3.3.4 Opening

One or more iterations of erosion are followed by the same number of iterations of dilation

Basic effects are:

1. Somewhat like erosion, however it is less destructive than erosion in general.
2. The effect of the operator is to preserve foreground regions that have a similar shape to the structuring element
3. While eliminating all other regions of foreground pixels.

3.3.5 Closing

One or more iterations of dilation followed by the same number of iterations of erosion.

The basic effects are:

1. boundaries can be smoothed
2. narrow gaps joined
3. small noise holes filled

CHAPTER 4

FEATURE SELECTION TECHNIQUES

To get the best results of recognition we have to apply the proper set of features. There are many techniques of feature selection. To the most popular belong principal component analysis, projection pursuit, correlation existing among features, correlation between the features and the classes, analysis of mean and variance of the features belonging to different classes, application of linear SVM feature ranking, etc. In this paper we have applied and compared some of them, including two linear SVM ranking methods, correlation analysis and the statistical analysis of clusters corresponding to the different classes.

4.1 THE SELECTION BASED ON THE MEAN AND VARIANCE OF THE DATA

The mean deviation of a data set focuses on the average of deviations of every observation x_i from the mean \bar{x} . As the sum of such deviations always equals zero by virtue of the definition of mean, the mean deviation is calculated using the absolute values of the deviations. Thus

Mean Deviation

$$\frac{\sum |x_i - \bar{x}|}{n}$$

Where $|x_i - \bar{x}|$ denotes the absolute value of each deviation (i.e.ignoring negative signs). This measure of variability is not widely used, preference being given to two related measures in which the offsetting effects of positive and negative deviations are eliminated by squaring. These measures are standard deviation and variance.

4.4.1 Standard Deviation

After summing the squared deviations it might seem obvious that the next step is to divide by n as with the mean deviation. this only applies when we are dealing

with a statistical population as opposed to a sample. A sample standard deviation is often used as an estimator of the population standard deviation and is obtained by dividing the sum of the squared deviations by $n-1$. Thus

Sample Standard Deviation

$$S = \sqrt{\left[\frac{\sum (x_i - \bar{x})^2}{n-1} \right]}$$

where \bar{x} the sample is mean and n is sample size

Population Standard Deviation

$$\sigma = \sqrt{\left[\frac{\sum (x_i - \mu)^2}{N} \right]}$$

where μ is the population mean and N is the population size

4.4.2 Variance

The variance is simply the standard deviation squared. Thus the sample variance is given by

Sample variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

and the population variance is given by

Population Variance

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Typically, the standard deviation is used in preference to the variance as a measure of dispersion because the units of measurement in variance are squared units and hence are difficult to interpret.

4.4.3 Coefficient of Variation

The coefficient of variation is a measure of relative dispersion. It is appropriate for comparing the variability within different data sets where the means of two data sets are different or where the units of measurement are different. It is defined as the standard deviation of the data sets as a percentage of its mean and, like the other measures of variability, it can be calculated for samples and populations.

Coefficient of variation

$$CV = \frac{s}{x} * 100$$

The most often used criterion of feature selection is the analysis of variance and means of the data samples belonging to each class. The variance of the feature describing the cells, belonging to one class should be as small as possible. Moreover to distinguish between different classes, the positions of means of feature values for the data belonging to different classes should be separated as much as possible.

Therefore the class oriented features should be considered to get the optimal choice of them. The multiclass problem should be solved by separating the task into two-class recognition sub-problems, as it is done in SVM classifiers (one against one mode of operation). The systematic policy for feature selection is to combine the variance and mean together to form single quality measure. He have done it by defining so called discrimination coefficient $SAB(f)$. For two classes A and B the discrimination coefficient of the feature f was defined as follows

$$S_{AB}(f) = \frac{|C_A(f) - C_B(f)|}{\sigma_A(f) + \sigma_B(f)}$$

C_A and C_B are the mean values of the feature f in the class A and B

σ_A and σ_B represent the standard deviations determined for both classes.

4.2 CORRELATION ANALYSIS

Correlation analysis is a means of measuring the strength or ‘closeness’ of the relationship between two variables. It should be clear that the concept of correlation is very closely linked to regression analysis. If all the paired points (x_i, y_i) lie on a straight line then the correlation between the variables x and y is perfect. Whether the correlation is perfectly positive or negative depends, of course, on whether the straight line through the points has a positive or negative slope.

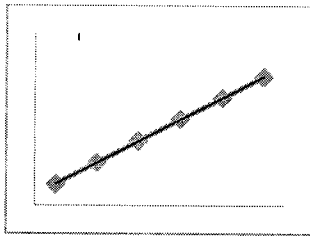


Fig.4.1 Perfect positive correlation

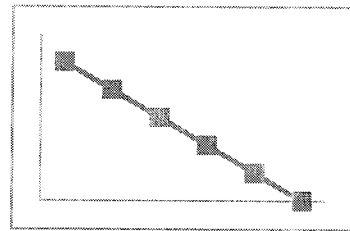


Fig.4.2: Perfect negative correlation

Correlation analysis provides a numerical summary measure of the degree of correlation between two variables x and y – a correlation coefficient denoted by r . this is defined so that its value must be within the range from -1 to $+1$ so that

$r=+1$ denotes perfect positive correlation

$r = 0$ denotes o correlation

$r= -1$ denotes perfect negative correlation

The correlation coefficient is defined and calculated as follow

Determination of correlation coefficient, r

$$r = \frac{S_{xy}}{S_x S_y}$$

where S_{xy} is the sample covariance, given by

$$\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

and S_x is the sample standard deviation of x values and S_y is the sample standard deviation of y values.

Computing the value of r using this expression is tedious. However, it reduces to a quick, convenient formula, expressed as follows

Computational formula for correlation coefficient, r

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}$$

The discriminative power of the candidate feature f for the recognition of the particular class among K classes can be also measured by the correlation of this feature with the class. Let us assume that the target class k is one among the classes forming target vector \mathbf{d} . Let us assume that the feature f is described by its unconditional and conditional means $m_c E\{f\} =$ and $m_{ck} E\{f|k\} =$. Assume that the variance $var(f)$ of feature f is known. The correlation between f and \mathbf{d} is derived from the covariance vector $\mathbf{cov}(f, \mathbf{d})$, related by the respective variance. The discriminative power of feature f is measured as the squared magnitude of the vector $\mathbf{corr}(f, \mathbf{d})$

$$S(f) = \frac{\sum_{k=1}^k P_k (m_{ck} - m_c)^2}{var(f \sum_{k=1}^k P_k (1 - P_k))}$$

Using this measure we can arrange the features in decreasing order from the highest to the smallest discriminative value.

CHAPTER 5

SUPPORT VECTOR MACHINE

5.1 Introduction

Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon, and Vapnik in COLT-92. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. In another terms, Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding overfit to the data. Support Vector machines can be defined as systems which use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory.

Support vector machine was initially popular with the NIPS community and now is an active part of the machine learning research around the world. SVM becomes famous when, using pixel maps as input; it gives accuracy comparable to sophisticated neural networks with elaborated features in a handwriting recognition task. It is also being used for many applications, such as hand writing analysis, face analysis and so forth, especially for pattern classification and regression based applications. The foundations of Support Vector Machines (SVM) have been developed by Vapnik and gained popularity due to many promising features such as better empirical performance.

The formulation uses the Structural Risk Minimization (SRM) principle, which has been shown to be superior, [4], to traditional Empirical Risk Minimization (ERM) principle, used by conventional neural networks. SRM minimizes an upper bound on the expected risk, where as ERM minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning. SVMs were developed to solve the classification problem, but recently they have been extended to solve regression problems

Machine Learning is considered as a subfield of Artificial Intelligence and it is concerned with the development of techniques and methods which enable the computer to learn. In simple terms development of algorithms which enable the machine to learn and perform tasks and activities. Machine learning overlaps with statistics in many ways. Over the period of time many techniques and methodologies were developed for machine learning tasks.

5.2 Statistical Learning Theory

The statistical learning theory provides a framework for studying the problem of gaining knowledge, making predictions, making decisions from a set of data. In simple terms, it enables the choosing of the hyper plane space such a way that it closely represents the underlying function in the target space .

In statistical learning theory the problem of supervised learning is formulated as follows. We are given a set of training data $\{(x_1, y_1) \dots (x_l, y_l)\}$ in $R^n \times R$ sampled according to unknown probability distribution $P(x, y)$, and a loss function $V(y, f(x))$ that measures the error, for a given x , $f(x)$ is "predicted" instead of the actual value y . The problem consists in finding a function f that minimizes the expectation of the error on new data that is, finding a function f that minimizes the expected error:

$$\int V(y, f(x)) P(x, y) dx dy$$

In statistical modeling we would choose a model from the hypothesis space, which is closest (with respect to some error measure) to the underlying function in the target space. More on statistical learning theory can be found on introduction to statistical learning theory

Why SVM?

Firstly working with neural networks for supervised and unsupervised learning showed good results while used for such learning applications. MLP's uses feed forward and recurrent networks. Multilayer perceptron (MLP) properties include universal approximation of continuous nonlinear functions and include learning with

input-output patterns and also involve advanced network architectures with multiple inputs and outputs.

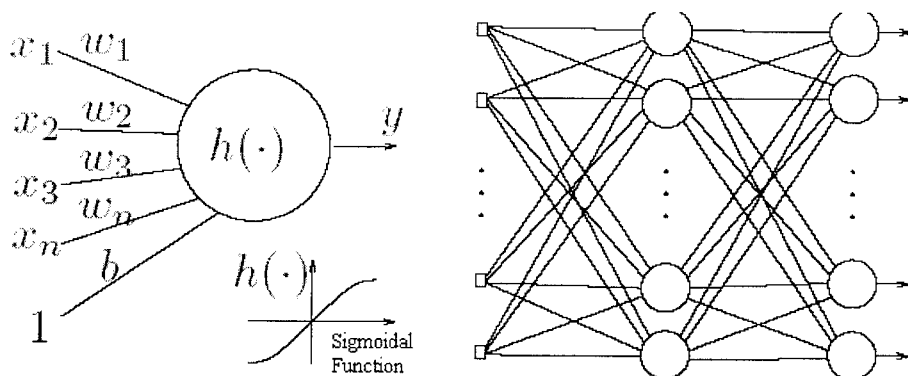


Fig. 5.1 : Simple Neural Network Multilayer Perceptron

There can be some issues noticed. Some of them are having many local minima and also finding how many neurons might be needed for a task is another issue which determines whether optimality of that NN is reached. Another thing to note is that even if the neural network solutions used tends to converge, this may not result in a unique solution. Now let us look at another example where we plot the data and try to classify it and we see that there are many hyper planes which can classify it. But which one is better?

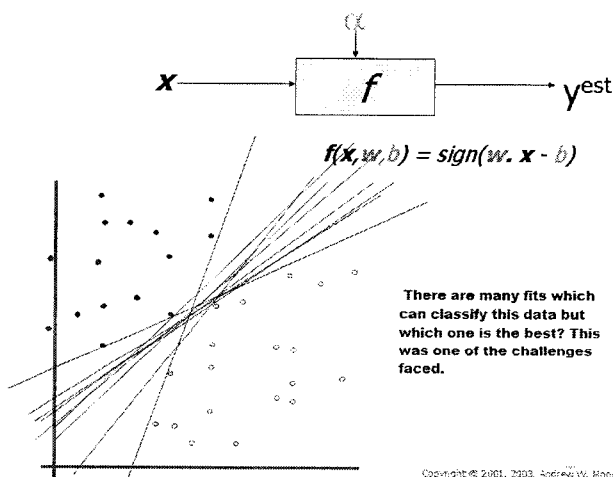


Fig.5.2 : Here we see that there are many hyper planes which can be fit in to classify the data but which one is the best is the right or correct solution. The need for SVM arises.

From above illustration, there are many linear classifiers (hyper planes) that separate the data. However only one of these achieves maximum separation. The reason we need it is because if we use a hyper plane to classify, it might end up closer to one set of datasets compared to others and we do not want this to happen and thus we see that the concept of maximum margin classifier or hyper plane as an apparent solution. The next illustration gives the maximum margin classifier example which provides a solution to the above mentioned problem.

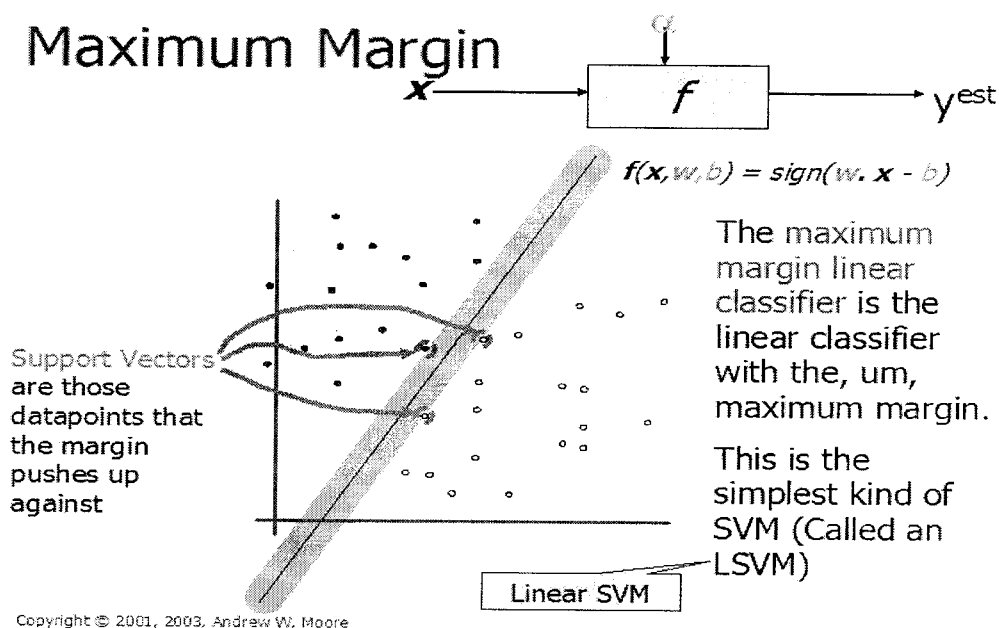


Fig 5.3: Illustration of Linear SVM.

Expression for Maximum margin is given as

$$\text{margin} \equiv \arg \min_{x \in D} d(x) = \arg \min_{x \in D} \frac{|x \cdot w + b|}{\sqrt{\sum_{i=1}^d w_i^2}}$$

The above illustration is the maximum linear classifier with the maximum range. In this context it is an example of a simple linear SVM classifier. Another interesting question is why maximum margin? There are some good explanations which include better empirical performance. Another reason is that even if we've made a small error in the location of the boundary this gives us least chance of causing a misclassification. The other advantage would be avoiding local minima and better classification. Now we try to express the SVM mathematically and for this

tutorial we try to present a linear SVM. The goals of SVM are separating the data with hyper plane and extend this to non-linear boundaries using kernel trick [8] [11]. For calculating the SVM we see that the goal is to correctly classify all the data. For mathematical calculations we have,

- [a] If $Y_i = +1$; $w x_i + b \geq 1$
- [b] If $Y_i = -1$; $w x_i + b \leq -1$
- [c] For all i ; $y_i (w x_i + b) \geq 1$

In this equation x is a vector point and w is weight and is also a vector. So to separate the data [a] should always be greater than zero. Among all possible hyper planes, SVM selects the one where the distance of hyper plane is as large as possible. If the training data is good and every test vector is located in radius r from training vector. Now if the chosen hyper plane is located at the farthest possible from the data. This desired hyper plane which maximizes the margin also bisects the lines between closest points on convex hull of the two datasets.

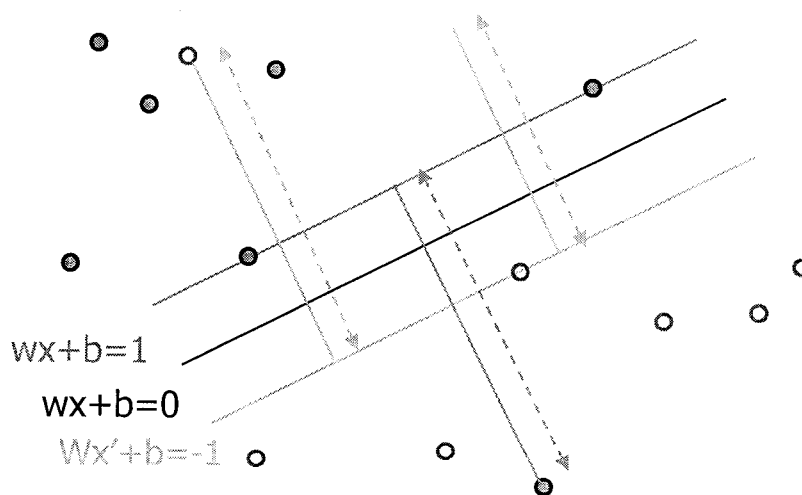


Fig.5.4: Representation of Hyper planes.

Distance of closest point on hyperplane to origin can be found by maximizing the x as x is on the hyper plane. Similarly for the other side points we have a similar scenario. Thus solving and subtracting the two distances we get the summed distance from the separating hyperplane to nearest points. Maximum Margin = $M = 2 / \|w\|$

Now maximizing the margin is same as minimum. Now we have a quadratic optimization problem and we need to solve for w and b . To solve this we need to optimize the quadratic function with linear constraints. The solution involves constructing a dual problem and where a Lagrangian's multiplier α_i is associated. We need to find w and b such that $\Phi(w) = \frac{1}{2} \|w'\|^2$ is minimized;

$$\text{And for all } \{(x_i, y_i)\}: y_i (w \cdot x_i + b) \geq 1.$$

Now solving: we get that $w = \sum \alpha_i \cdot x_i$, $b = y_k - w \cdot x_k$ for any x_k such that $\alpha_k \neq 0$

Now the classifying function will have the following form: $f(x) = \sum \alpha_i y_i x_i \cdot x + b$

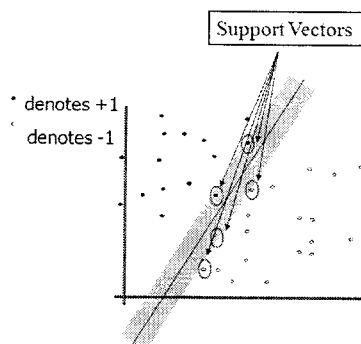


Fig. 5.5: Representation of Support Vectors

5.3 SVM for Classification

SVM is a useful technique for data classification. Even though it's considered that Neural Networks are easier to use than this, however, sometimes unsatisfactory results are obtained. A classification task usually involves with training and testing data which consist of some data instances. Each instance in the training set contains one target values and several attributes. The goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only the attributes.

Classification in SVM is an example of Supervised Learning. Known labels help indicate whether the system is performing in a right way or not. This information points to a desired response, validating the accuracy of the system, or be used to help the system learn to act correctly. A step in SVM classification involves identification as which are intimately connected to the known classes. This is called feature selection or feature extraction. Feature selection and SVM classification together

have a use even when prediction of unknown samples is not necessary. They can be used to identify key sets which are involved in whatever processes distinguish the classes

5.4 Properties of SVM

1. Flexibility in choosing a similarity function
2. Sparseness of solution when dealing with large data sets
 - only support vectors are used to specify the separating hyperplane
3. Ability to handle large feature spaces
 - complexity does not depend on the dimensionality of the feature space
4. Over fitting can be controlled by soft margin approach
5. Nice math property: a simple convex optimization problem which is guaranteed to converge to a single global solution
6. Feature Selection

5.5 SVM Applications

SVM has been used successfully in many real-world problems

1. Text (and hypertext) categorization
2. Image classification
3. Bioinformatics (Protein classification, Cancer classification)
4. Hand-written character recognition
5. Machine vision: e.g face identification
 - Outperforms alternative approaches (1.5% error)
6. Handwritten digit recognition: USPS data
 - Comparable to best alternative (0.8% error)

7. Can modify SVM technique for numeric prediction problems

5.6 Weakness of SVM

1. It is sensitive to noise
 - A relatively small number of mislabeled examples can dramatically decrease the performance
2. It only considers two classes

How to do multi-class classification with SVM?

1. With output parity m , learn m SVM's
 - SVM 1 learns "Output==1" vs "Output != 1"
 - SVM 2 learns "Output==2" vs "Output != 2"
 - :
 - SVM m learns "Output== m " vs "Output != m "
2. To predict the output for a new input, just predict with each SVM and find out which one puts the prediction the furthest into the positive region.

CHAPTER 6

EXPERIMENTAL RESULTS AND DISCUSSION

The experimental results are shown below. The first process is the conversion of RGB image into Gray scale image which is then fed for cell extraction

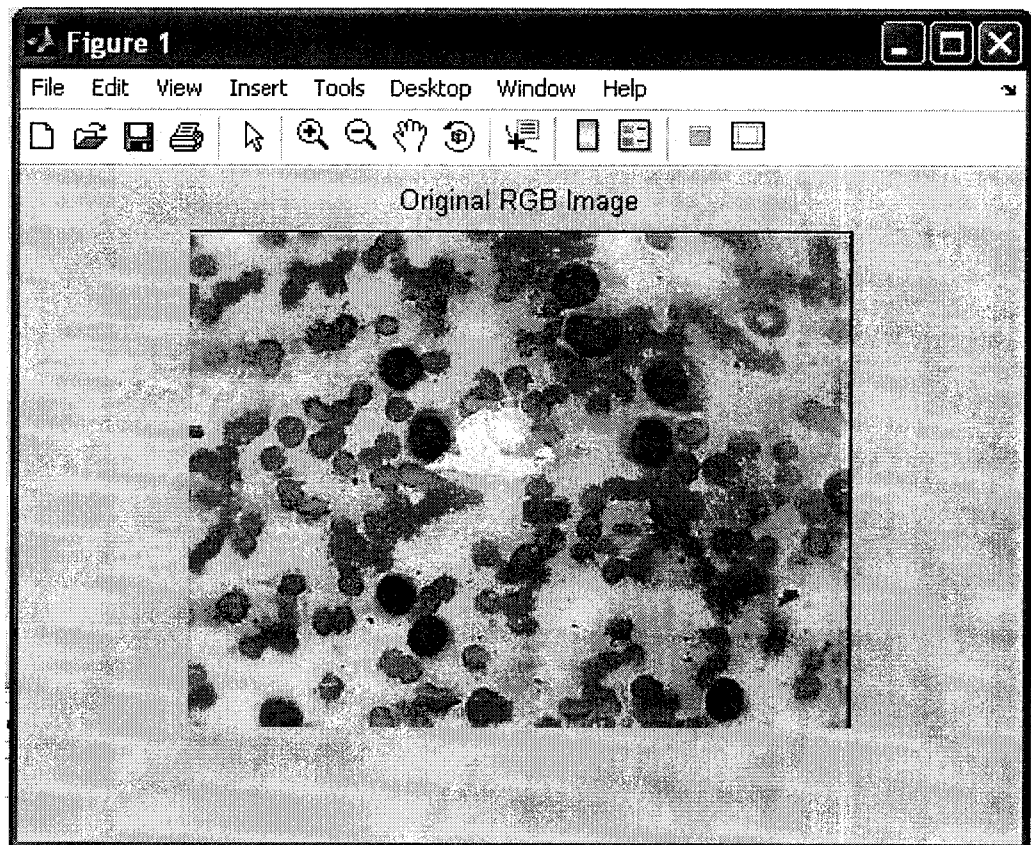


Fig 6.1 : Original RGB Image fed for Segmentation

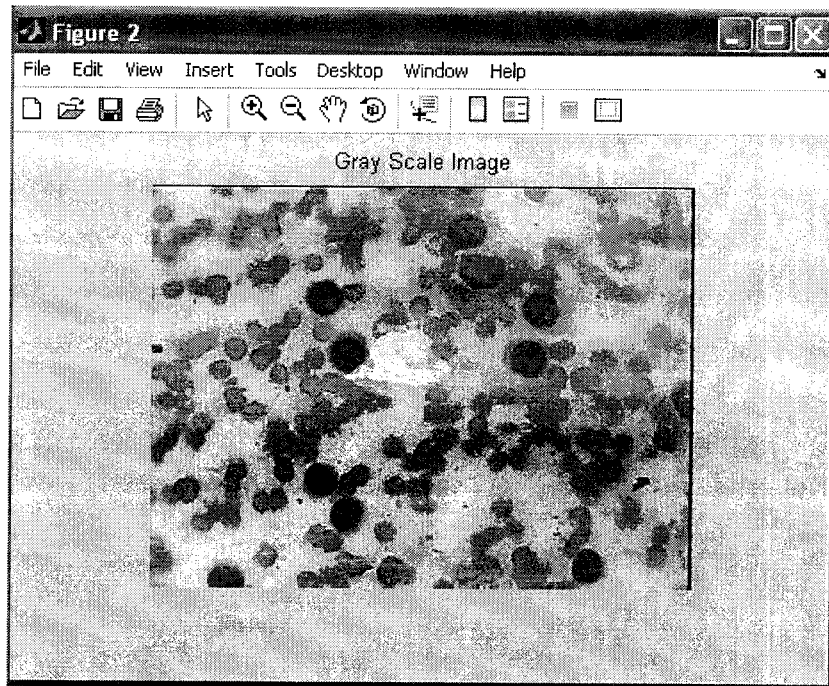


Fig 6.2: Converted Gray Scale Image

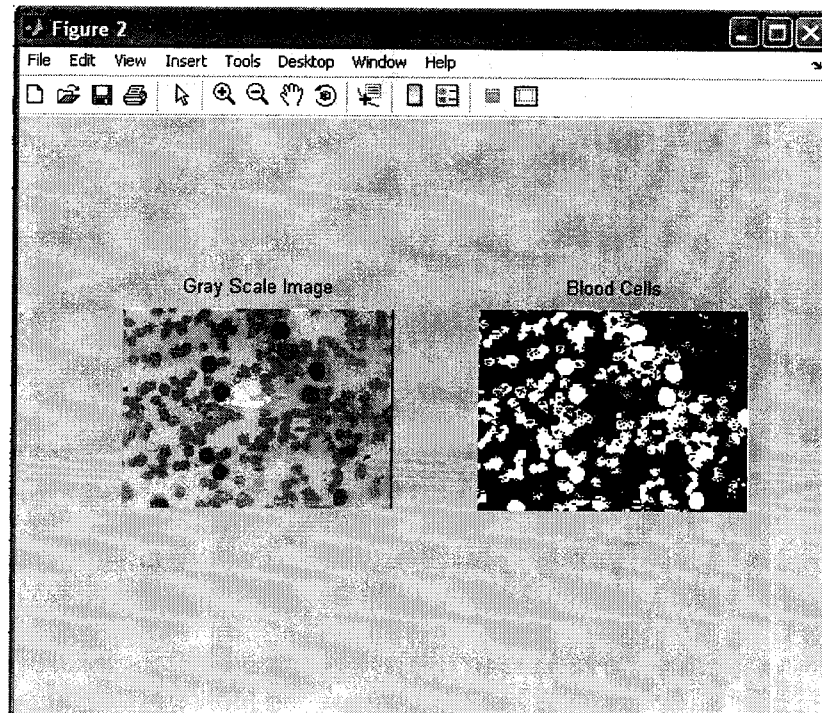


Fig 6.3: Cells obtained after Morphological Operations

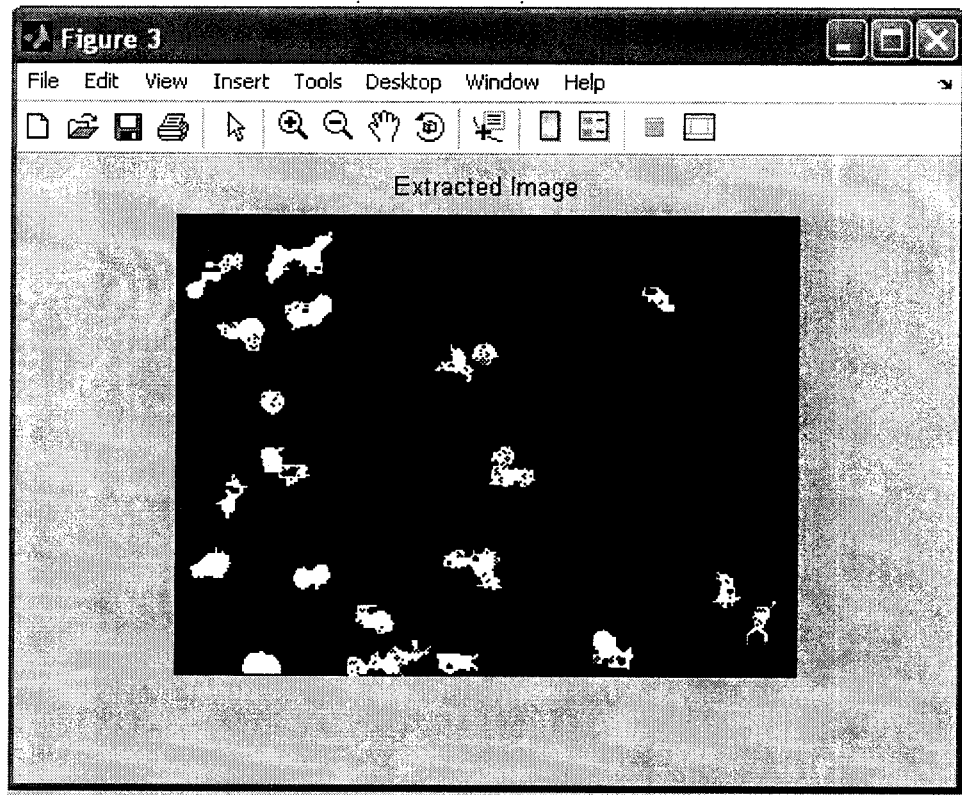


Fig 6.4: Extracted Image

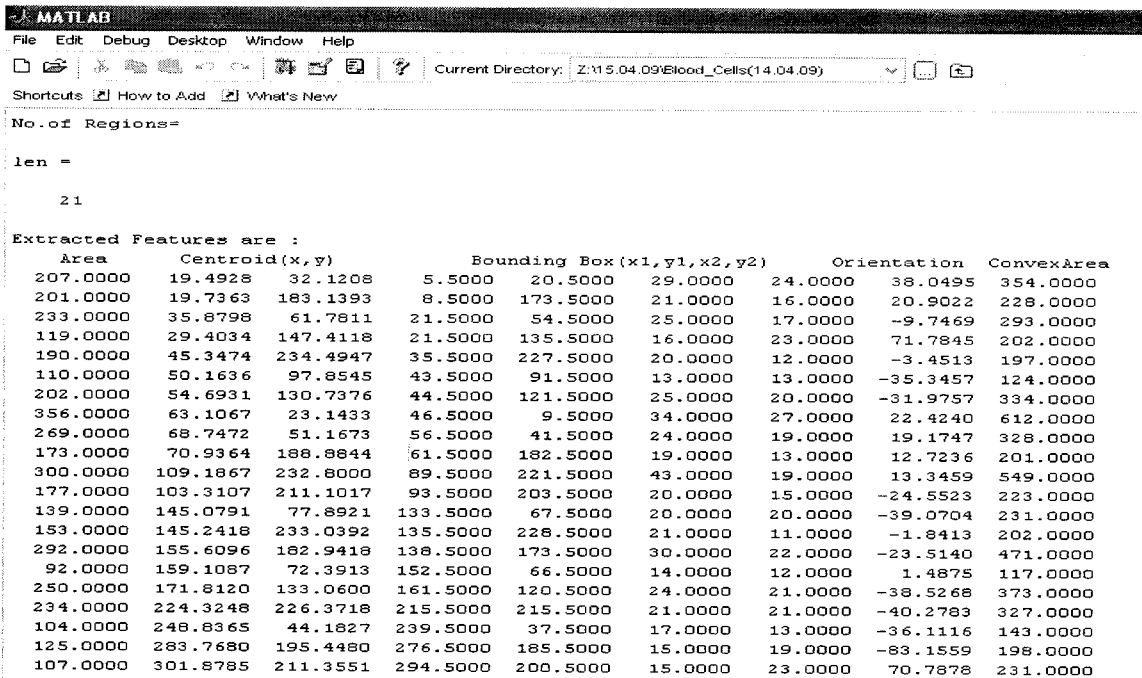


Fig 6.5: The Extracted Features of individual Cell

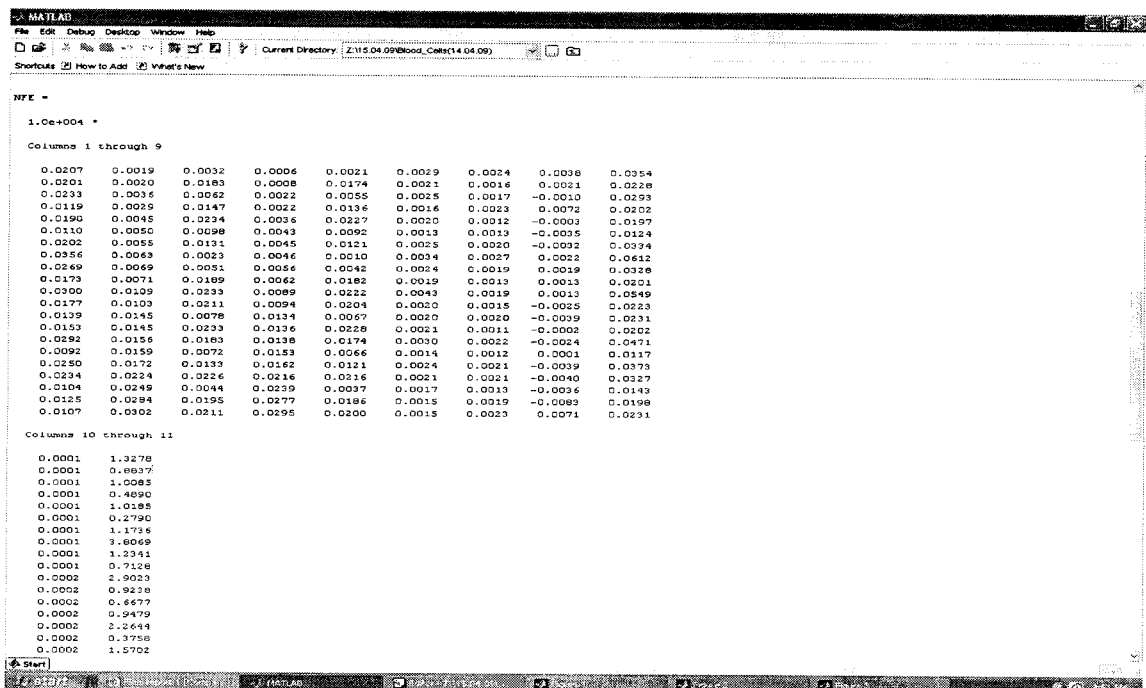


Fig 6.6: Non- Ranked Features and calculation of Mean and Variance

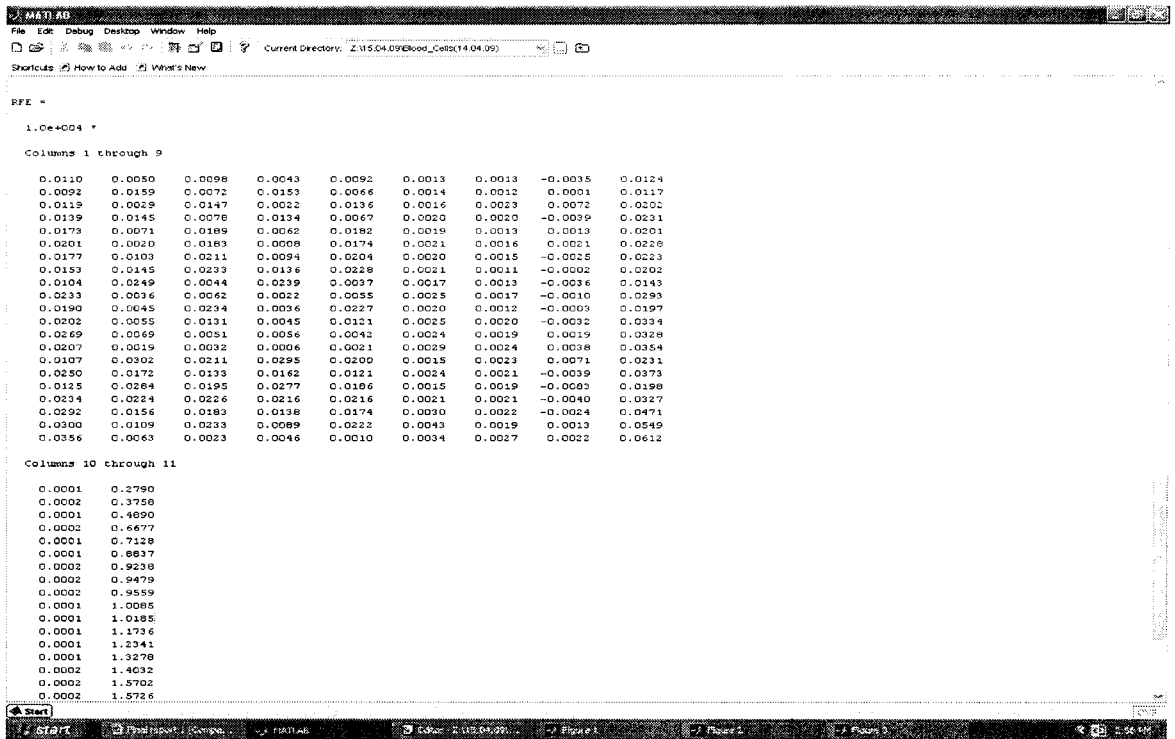


Fig 6.7: Ranked Features for Features Selection

Threshold: Cells ranging between 100 and 200

No of cells Extracted: 8

No of Samples to Train	Error Rate		
	Variance	SVM	Correlation
1	0.3750	0.3750	0.7826
3	0.2500	0.3750	0.7500
5	0.1250	0.2500	0.3750
6	0.1250	0.2500	0.2500
7	0.1125	0.1250	0.1230

Table :6.1 Representation of cells between 100 & 200

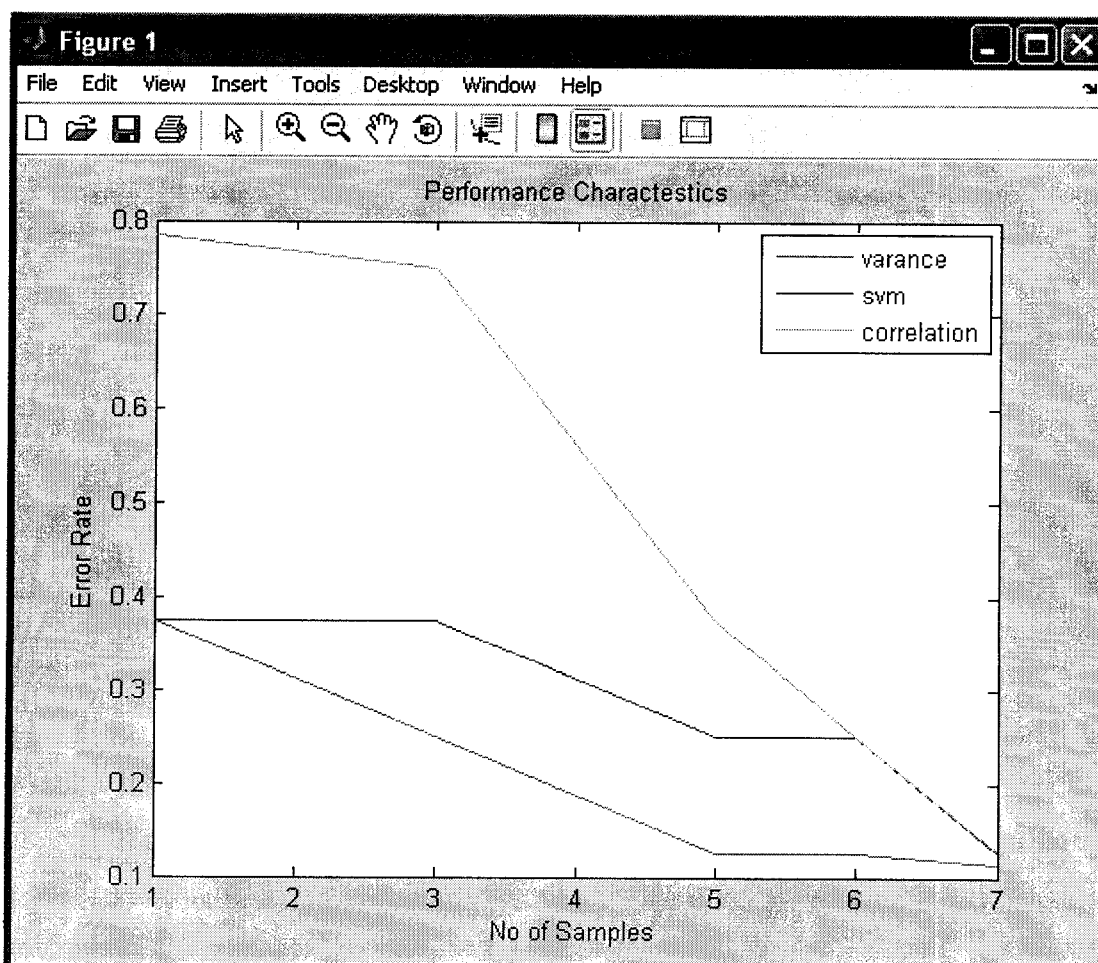


Fig 6.8 performance characteristics of cell between 100 & 200

Threshold: Cells ranging between 100 and 250

No of cells Extracted: 12

No of Samples to Train	Error Rate		
	Variance	SVM	Correlation
1	0.4167	0.4167	0.6750
3	0.3333	0.3333	0.6667
5	0.2500	0.3333	0.6625
7	0.1667	0.2500	0.4167
9	0.0833	0.2500	0.2500

Table :6.2 Representation of cells between 100 & 250

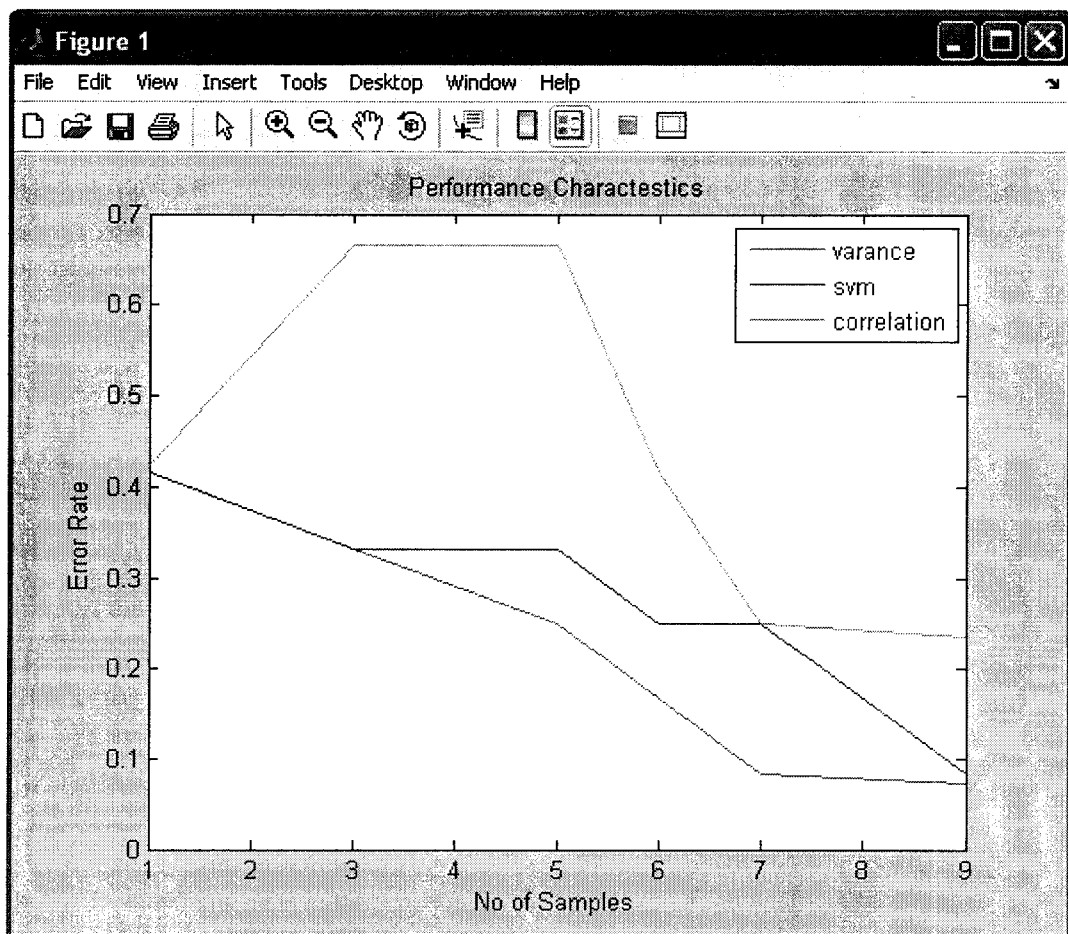


Fig 6.9 performance characteristics of cell between 100 & 250

Threshold: Cells ranging between 100 and 400

No of cells Extracted: 17

No of Samples to Train	Error Rate		
	Variance	SVM	Correlation
1	0.4706	0.4706	0.6725
3	0.4188	0.4706	0.5882
5	0.3529	0.4188	0.5882
8	0.2941	0.2941	0.5882
12	0.2353	0.1765	0.2941
15	0.0588	0.05888	0.1176

Table :6.3 Representation of cells between 100 & 400

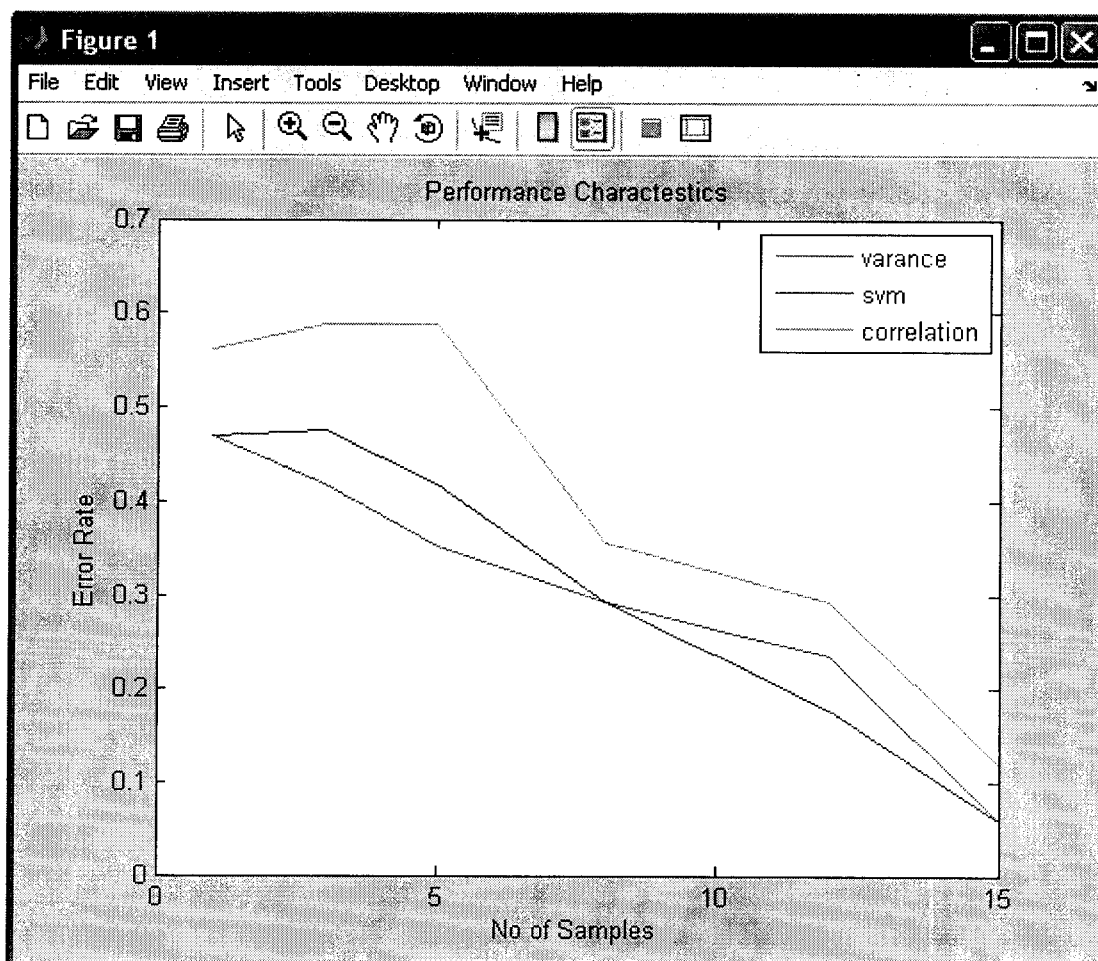


Fig 6.10 performance characteristics of cell between 100 & 400

Threshold: Cells ranging between 10 and 400

No of cells Extracted: 43

No of Samples to Train	Error Rate		
	Variance	SVM	Correlation
5	0.5185	0.4815	0.5556
10	0.3333	0.3333	0.5556
15	0.3333	0.2593	0.4444
20	0.2593	0.1852	0.2593
25	0.0370	0.0370	0.0741

Table :6.1 Representation of cells between 10 & 400

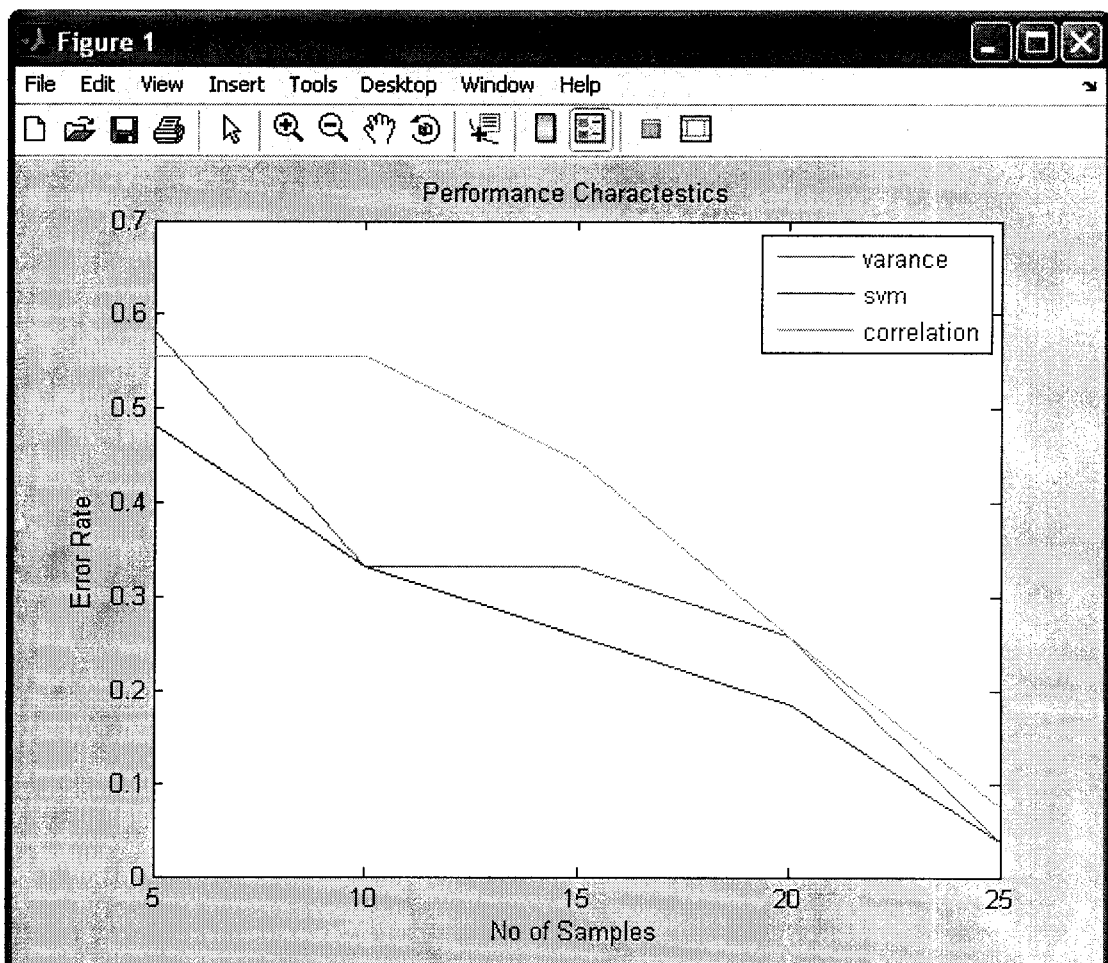


Fig 6.11 performance characteristics of cell between 10 & 400

CHAPTER 7

CONCLUSION AND FUTURE WORK

The extraction of individual cells plays an important role in feature generation as well as feature selection. The results of experiments prove the important role of feature selection. Irrespective of the applied method, the selection improves significantly the accuracy of recognition. Among the best feature selection methods is the correlation between the feature and the class as well as the linear SVM ranking based on single feature application.

Experimental results have stated that the recognition error rate obtained from variance, Correlation and SVM classification method goes on decreasing when the number of samples of each cell image increases

Our future work involves obtaining better recognition error rate by using less number of samples. Feature generation of less number of cells extracted will produce better results and will predict the percentage of affected cells for leukemia disease

APPENDICES

APPENDIX I

Matlab code for cell image segmentation and Feature Generation

```
clc;
clear all;
close all;
addpath(genpath(pwd));
%-----
% Learning and Learning Parameters
nbclass=2;
c = 1000;
lambda = 1e-7;
kernel option= 1;
kernel='gaussian';
verbose = 0;
%%%Classification Samples
nos=14; %%no.of samples to train
% Testp=50;
sigma=0.5;
%x=input('Enter the file name ','s'); %RGB Image
%I=imread(x);
k=input('Enter the Option=')
option=k; %%Feature Selection Parameters
% % % nbclass=2;
% % % xxx
I=imread('bone.jpg');
imshow(I)
I=rgb2gray(I);

figure;
subplot(1,2,1);
imshow(I);
title('Original Image');

% SI=watershed(I);
% SI=I<200;
SI=im2bw(I);
SI=~SI;
subplot(1,2,2);
imshow(SI);
title('Blood Cells');
figure;
% % % se = strel('disk',3);
```

```

% % % EI=imerode(SI,se);
% % % DI=imdilate(EI,se);
% % % OI=bwmorph(EI,'open',Inf);
% % % CI=bwmorph(EI,'close',Inf);
bw=SI;
bw2=bwmorph(bw,'thin');
% figure,imshow(bw2);
bw2=bwmorph(bw2,'fill','Inf');
% figure,imshow(bw2);

bw2=bwmorph(bw2,'diag','Inf');
% figure,imshow(bw2);
bw2=bwmorph(bw2,'bridge','Inf');
% figure,imshow(bw2);
L=bwlabeln(bw2);
S=regionprops(L,'FilledArea');
BW2=ismember(L,find([S.FilledArea]<400));
% figure,imshow(BW2);
bw2=bwmorph(BW2,'clean','Inf');
% figure,imshow(bw2);
L=bwlabeln(bw2);
S=regionprops(L,'FilledArea');
BW2=ismember(L,find([S.FilledArea]>100));
% figure,imshow(BW2);

CI=BW2;
imshow(CI);
title('Blood Cells');
LI=bwlabel(CI);
% RGB = label2rgb(LI);
% imshow(RGB);

title('Labelled Image');
F=regionprops(LI,'All');
if (isempty(F))
    disp('No Features to Extract...');
    msgbox('No Features to Extract...','Feature
Extraction');
    return;
end;
disp('No.of Regions=');
len=size(F,1)
FE=[];
for i=1:len
    if length(F(i).Image) >10
%         figure;
%         imshow(F(i).Image);

```

```

        f=[F(i).Area F(i).Centroid F(i).BoundingBox
F(i).Orientation F(i).ConvexArea];
        FE=[FE;f];
    end
end
disp('Extracted Features are :');
disp('      Area      Centroid(x,y)      Bounding
Box(x1,y1,x2,y2)      Orientation ConvexArea');
disp(FE);

%%Add classlabels
xapp=FE;
[hh ww]=size(FE);
% trainidx=hh*trainp;
% xtest=xapp;
% xapp=xapp(1:trainidx,:);
[hh ww]=size(FE);
yapp(1:hh)=1;
yapp(hh/2: hh)=2;
yapp=yapp';

FE(:,ww+1)=yapp;

%%%Feature Selection Starts
%%%mean=1;

NFE=FE;
[h w]=size(FE);
for i=1:h
    if option==1 %%% Variance
        NFE(i,w+1)=var(NFE(i,:)); %%%Add Features at
last Columns
    elseif option==2 %%% Correlation
        temps=xcorr(NFE(i,:)); %%%Add Features at
last Columns
        NFE(i,w+1)=temps(1);
    end
end
end

%%%Sort Rows Based on teh Ranks
[h w]=size(NFE);
%%%RFE-Ranked Fetatures based on the Ranks (ie) Last
Columns
RFE=sortrows(NFE,w);
disp('Selected Features are :');
if option<=2
SFE=RFE(:,1:w-2); %%%Remove Ranking Column

```

```
yappclass=RFE(:,w-1); %%%Class Labels
else
SFE=RFE(:,1:w-1); %%%no need to remove Column
yappclass=RFE(:,w); %%%Class Labels
end
SFE
NFE
RFE

xapp=SFE(1:nos,:);
yapp=yappclass(1:nos);
[n1, n2]=size(xapp);
yapp(size(yapp,1))=nbclass;
```

APPENDIX II

Matlab code for SVM Classification

```
kerneloptionm.matrix=svkernel(xapp, kernel, kerneloption);
if(option==1 |option==2)

[xsup,w,b,nbsv,classifier,pos]=svmmulticlassoneagainstone
([],yapp,nbclass,c,lambda,'numerical',kerneloptionm,verbo
se);
else

[xsup,w,b,nbsv]=svmmulticlassoneagainstall(xapp,yapp,nbcl
ass,c,lambda, kernel, kerneloption, verbose);
end

xtest=SFE;
% [ypred,maxi] =
svmmultivaloneagainstone(xtest,zsup,w,b,nbsv, kernel, kerne
loption);

if option==1 |option==2
kerneloptionm.matrix=svkernel(xtest, kernel, kerneloption,
xapp(pos,:));
[ypred,maxi] =
svmmultivaloneagainstone([], [],w,b,nbsv,'numerical',kerne
loptionm);
else
[ypred,maxi] =
svmmultival(xtest,xsup,w,b,nbsv, kernel, kerneloption);
end

cp = classperf(yappclass);
classperf(cp,ypred,logical(ypred));
disp('Error Rate :');
disp(cp.errorRate);
disp('If u need more metrics type <--cp--> in teh
prompt:');
```

REFERENCES

1. S. Osowski, T. Markiewicz, B. Mariańska, L. Moszczyński, Feature generation for the cell image recognition of myelogenous leukemia, IEEE Int. Conf. EUSIPCO, Vienna, pp. 753-756, 2007.
2. N. Theera-Umpon, P. Gader, system-level training of neural networks for counting white blood cells, IEEE Trans. SMS-C, 32:48-53, 2006
3. L. Shafarenko, M. Petrou, and J. Kittler, "Automatic watershed segmentation of textured color images," *IEEE Trans. Image Processing*, vol. 6, pp. 1530–1543, Nov. 2006.
4. S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using a support vector machine. In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, *Neural Networks for Signal Processing VII — Proceedings of the 2007 IEEE Workshop*, New York, 2007. IEEE.
5. F. Meyer, "Color image segmentation," in *Proc. Int. Conf. Image Processing Applications*, 2006, pp. 303–306.
6. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2007.