

3-24-2011

Automated Detection of Hematological Abnormalities through Classification of Flow Cytometric Data Patterns

Mark A. Rossman

Florida International University, locusman_fl@yahoo.com

Follow this and additional works at: <http://digitalcommons.fiu.edu/etd>



Part of the [Biomedical Commons](#), and the [Signal Processing Commons](#)

Recommended Citation

Rossman, Mark A., "Automated Detection of Hematological Abnormalities through Classification of Flow Cytometric Data Patterns" (2011). *FIU Electronic Theses and Dissertations*. 344.
<http://digitalcommons.fiu.edu/etd/344>

This work is brought to you for free and open access by the University Graduate School at FIU Digital Commons. It has been accepted for inclusion in FIU Electronic Theses and Dissertations by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

AUTOMATED DETECTION OF HEMATOLOGICAL ABNORMALITIES
THROUGH CLASSIFICATION OF FLOW CYTOMETRIC DATA PATTERNS

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ELECTRICAL ENGINEERING

by

Mark Alexander Rossman

2011

To: Dean Amir Mirmiran
College of Engineering and Computing

This dissertation, written by Mark Alexander Rossman, and entitled Automated Detection of Hematological Abnormalities through Classification of Flow Cytometric Data Patterns, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Nikolaos Tsoukias

Jean Andrian

Armando Barreto

Malek Adjouadi, Major Professor

Date of Defense: March 24, 2011

The dissertation of Mark Alexander Rossman is approved.

Dean Amir Mirmiran
College of Engineering and Computing

Interim Dean Kevin O'Shea
University Graduate School

Florida International University, 2011

DEDICATION

To my family and friends

ACKNOWLEDGMENTS

My deepest appreciation goes to my advisor, Dr. Malek Adjouadi, for the admirable support and guidance provided for the completion of this work, as well as for his incomparable dedication to mentoring.

My appreciation extends also to my committee members: Dr. Armando Barreto, Dr. Nikolaos Tsoukias, and Dr. Jean Andrian, for the time they spent advising and encouraging me throughout this project.

My gratitude also is given to the Beckman Coulter Corporation, for giving me access to the hematological data and software that allowed me to perform this research. I would like to thank my colleagues, especially John Riley and Melvin Ayala for their assistance on different tasks related to my research.

Last, but definitely not least, I appreciate the support provided by the National Science Foundation under grants (CNS-BPC-AE-1042341, CNS-MRI-R2-0959985, and HRD-CREST-0833093).

ABSTRACT OF THE DISSERTATION

AUTOMATED DETECTION OF HEMATOLOGICAL ABNORMALITIES
THROUGH CLASSIFICATION OF FLOW CYTOMETRIC DATA PATTERNS

by

Mark Alexander Rossman

Florida International University, 2011

Miami, Florida

Professor Malek Adjouadi, Major Professor

Flow Cytometry analyzers have become trusted companions due to their ability to perform fast and accurate analyses of human blood. The aim of these analyses is to determine the possible existence of abnormalities in the blood that have been correlated with serious disease states, such as infectious mononucleosis, leukemia, and various cancers. Though these analyzers provide important feedback, it is always desired to improve the accuracy of the results. This is evidenced by the occurrences of misclassifications reported by some users of these devices. It is advantageous to provide a pattern interpretation framework that is able to provide better classification ability than is currently available. Toward this end, the purpose of this dissertation was to establish a feature extraction and pattern classification framework capable of providing improved accuracy for detecting specific hematological abnormalities in flow cytometric blood data.

This involved extracting a unique and powerful set of shift-invariant statistical features from the multi-dimensional flow cytometry data and then using these features as inputs to a pattern classification engine composed of an artificial neural network (ANN). The

contribution of this method consisted of developing a descriptor matrix that can be used to reliably assess if a donor's blood pattern exhibits a clinically abnormal level of variant lymphocytes, which are blood cells that are potentially indicative of disorders such as leukemia and infectious mononucleosis.

This study showed that the set of shift-and-rotation-invariant statistical features extracted from the eigensystem of the flow cytometric data pattern performs better than other commonly-used features in this type of disease detection, exhibiting an accuracy of 80.7%, a sensitivity of 72.3%, and a specificity of 89.2%. This performance represents a major improvement for this type of hematological classifier, which has historically been plagued by poor performance, with accuracies as low as 60% in some cases.

This research ultimately shows that an improved feature space was developed that can deliver improved performance for the detection of variant lymphocytes in human blood, thus providing significant utility in the realm of suspect flagging algorithms for the detection of blood-related diseases.

TABLE OF CONTENTS

CHAPTER	PAGE
1. INTRODUCTION	1
2. BACKGROUND ON CLINICAL HEMATOLOGY AND BLOOD CELLS	6
2.1. Background and Significance of Blood Cells.....	6
2.1.1. Basic Categories of Human Blood Cells	6
2.1.2. Abnormal Variations in White Blood Cells.....	7
2.1.3. Benefits of the WBC Manual Differential	8
2.1.4. Limitations of the WBC Manual Differential Process.....	10
2.2. Automated Hematology	11
2.2.1. Automated Hematology Devices	11
2.2.2. VCS Automated Hematology Devices	12
2.2.3. Uses of Automated Hematology Analyzers As Screening Tools	12
2.2.4. Examples of Common Hematological Patterns	14
2.2.5. Limitations of Automated Hematology Analyzers.....	19
2.3. Overview of Methods for Detection of Abnormal Hematological Patterns	20
2.3.1. Basic Operational Steps of Existing Classification Methods	20
2.3.2. Method of Performance Evaluation of Detection Algorithms.....	22
3. FEATURE EXTRACTION METHODS FOR VARIANT LYMPHOCYTE CLASSIFICATION	27
3.1. Problem Introduction	27
3.2. Variant Lymphocyte Experimental Dataset.....	30
3.3. Case Studies of Typical Variant Lymphocyte Hematological Patterns.....	31
3.4. Selection of Pattern Features for Variant Lymphocyte Sample Classification	40
3.5. Practical Difficulties With Variant Lymphocyte Classification Methods	45
3.6. Theory of Fisher Linear Discriminants.....	48
3.7. Use of Fisher Linear Discriminants for Feature Selection and Evaluation	50
3.8. Multi-Dimensional Feature Extraction Method for Improved Pattern Recognition.....	51
3.8.1. Mathematical Aspects of the Principal Component Transform.....	52
3.8.2. Principal Component Transform for 3D Variance and Pose Representation	55
3.8.3. Extracting Eigensystem-Based Population Descriptors.....	57
3.8.4. Proposed Extraction Method for Eigensystem-Based Cluster Descriptors	58
3.8.5. Comparison of Features Using Fisher Linear Discriminant Analysis	59
3.8.6. Retrospective.....	63

4. APPLICATION OF MACHINE LEARNING TO THE DETECTION OF VARIANT LYMPHOCYTE DATA PATTERNS	66
4.1. Objectives	66
4.2. Experimental Setup.....	68
4.3. Dimensions of Feature Space	68
4.4. Appropriate Artificial Neural Network Design	69
4.5. Data Partitioning.....	70
4.6. Artificial Neural Network Design Considerations.....	71
4.6.1. Type of Neural Network	71
4.6.2. Learning Rule.....	71
4.6.3. Targets Used For Supervised Learning.....	71
4.6.4. Type and Number of Input Units	72
4.6.5. Activation Functions.....	72
4.6.6. Training Procedure.....	72
4.6.7. Learning Rates	73
4.6.8. Number of Hidden Units.....	73
4.7. Strategies to Find Optimal ANN Topology.....	73
4.7.1. Perceptron ANN Experiment #1	74
4.7.2. Perceptron ANN Experiment #2.....	76
4.7.3. Perceptron ANN Experiment #3.....	79
4.7.4. Perceptron ANN Experiment #4.....	81
4.7.5. Perceptron ANN Experiment #5.....	86
4.8. Discussion of Perceptron ANN Topology Experiments.....	90
4.9. Finalized Perceptron ANN Topology and Best Performance.....	91
5. CLASSIFICATION OF VARIANT LYMPHOCYTES USING FEATURE EXTRACTION AND MACHINE LEARNING	93
5.1. Objectives of the Method.....	93
5.2. Data Collection	93
5.3. Feature Extraction.....	94
5.3.1. ANN Configuration and Training Procedure.....	94
5.4. Testing Results.....	97
5.5. Discussion of Classification Performance Improvements Obtained	97
5.6. Samples Classified by Implemented Approach.....	98
5.6.1. Pattern #1: Variant Lymph Positive Sample Correctly Detected	98
5.6.2. Pattern #2: Variant Lymphocyte Negative Sample Incorrectly Detected.....	99
5.6.3. Pattern #3: Variant Lymph Positive Sample Correctly Detected	100
6. CONCLUSIONS.....	103
BIBLIOGRAPHY.....	106
VITA.....	110

LIST OF TABLES

TABLE	PAGE
Table 2.1: Example of a 200-Cell Manual WBC Differential Count Report	9
Table 2.2: Example of a 200-Cell Manual WBC Differential Percentage Report.....	9
Table 2.3: Breakdown of Hematological Malignancies in the United States.....	13
Table 2.4: Entries of a confusion matrix.....	22
Table 3.1: LH750 Differential Algorithm Reported WBC Percentages for Case #1	32
Table 3.2: Manual Differential WBC Percentages for Case #1	33
Table 3.3: LH750 Algorithm Reported WBC Percentages for Case #2.....	35
Table 3.4: Manual Differential WBC Percentages for Case #2.....	35
Table 3.5: LH750 Algorithm Reported WBC Percentages for Case #3	37
Table 3.6: Manual Differential WBC Percentages for Case #3.....	37
Table 3.7: LH750 Algorithm Reported WBC Percentages for Case #4.....	38
Table 3.8: Manual Differential WBC Percentages for Case #4.....	39
Table 3.9: ROC Performance Metrics of Common Lymphocyte Population Statistical Parameters.....	48
Table 3.10: Raw Event Array for a single hematological population collected with an LH750 analyzer.....	56
Table 3.11: Area-Under-The-Curve Values for the 10 Best Eigensystem Parameters.....	63
Table 4.1: Eigensystem Orientation and Variance Parameters for 4 WBC Populations	68
Table 4.2: Perceptron ANN Configuration Variables for Training Experiment #1	74
Table 4.3: Confusion Matrix and Performance Statistics for Best Trial of Experiment #1.....	75
Table 4.4: Sigmoid Activation Function Parameter Values	77
Table 4.5: Perceptron ANN Configuration Variables for Training Experiment #2	78

Table 4.6: Confusion Matrix and Performance Statistics for Best Trial of Experiment #2.....	78
Table 4.7: Perceptron ANN Configuration Variables for Training Experiment #3	80
Table 4.8: Data Partitions for Training, Cross-Validation, and Testing.....	82
Table 4.9: Performance Result Ranges for Experiment #5, Trial #1	87
Table 4.10: Confusion Matrix and Performance Statistics for Best Trial of Experiment #5, Trial #1	87
Table 4.11: Performance Result Ranges for Experiment #5, Trial #2.....	88
Table 4.12: Confusion Matrix and Performance Statistics for Best Trial of Experiment #5, Trial #2.....	88
Table 4.13: Performance Result Ranges for Experiment #5, Trial #3.....	89
Table 4.14: Confusion Matrix and Performance Statistics for Best Trial of Experiment #5, Trial #3	89
Table 4.15: Average Performance Metrics for 3 Topologies Tested in ANN Experiment #5.....	91
Table 4.16: Perceptron ANN Configuration Values of Optimal Network	91
Table 4.17: Performance of LH750 Variant Lymph Suspect Flag on Testing + Cross-Validation Datasets	92
Table 4.18: Performance of Proposed Variant Lymph Suspect Flag on Testing + Cross-Validation Datasets	92
Table 5.1: Donor Information and Data Used	94
Table 5.2: LH750 Variant Lymphocyte Suspect Flag Performance.....	97
Table 5.3: Eigensystem-Based Variant Lymphocyte Suspect Flag Performance.....	97
Table 5.4: Manual WBC Differential for Discussion Pattern #1	99
Table 5.5: Manual WBC Differential for Discussion Pattern #2.....	100
Table 5.6: Manual WBC Differential for Discussion Pattern #3.....	101

LIST OF FIGURES

FIGURE	PAGE
Figure 2.1: Basic WBC Types: (a) Lymphocyte, (b) Monocyte, (c) Neutrophil, (d) Eosinophil, (e) Basophil	7
Figure 2.2: VCS WBC Differential Pattern of Normal Blood Donor – Before Event Categorization	14
Figure 2.3: VCS WBC Differential Pattern of Normal Blood Donor – After Event Categorization.....	15
Figure 2.4: WBC Differential Patterns of Donor Diagnosed with Infectious Mononucleosis.....	16
Figure 2.5: WBC Differential Patterns of Donor Diagnosed with Chronic Lymphocytic Leukemia	17
Figure 2.6: WBC Differential Patterns of Donor Diagnosed with Acute Lymphocytic Leukemia	18
Figure 2.7: WBC Patterns of Donor Diagnosed with Acute Myelocytic Leukemia	19
Figure 2.8: Two ROC curves plotted from the confusion matrixes of two classifiers	26
Figure 3.1 VCS WBC Differential Pattern of Normal Blood Donor.....	28
Figure 3.2 VCS WBC Differential Blood Pattern of Blood Donor with Lymphocytic Leukemia	28
Figure 3.3: Normal (left) and Lymphocytic Leukemia (right) Lymphocyte Populations	29
Figure 3.4: Blood Pattern with Abnormal Level of Variant Lymphocytes – Case #1	32
Figure 3.5: VCS WBC Population Statistical Parameters – Case #1	33
Figure 3.6: VCS WBC Population Statistical Parameters - Normal Blood Sample.....	33
Figure 3.7: Percentage Difference of Statistical Parameters (Variant Lymph Sample minus Normal Sample).....	34
Figure 3.8: Blood Pattern with Abnormal Level of Variant Lymphocytes – Case #2	35

Figure 3.9: Blood Pattern with Normal Level of Variant Lymphocytes – Case #3.....	36
Figure 3.10: Blood Pattern with Blasts but Normal Level of Variant Lymphocytes – Case #4.....	37
Figure 3.11: Histogram Overlay of Lymphocyte Volume SD.....	41
Figure 3.12: Histogram Overlays of Lymphocyte Population Statistical Parameters.....	42
Figure 3.13: Histogram Overlay with Sliding Decision Point Bar.....	43
Figure 3.14: ROC Curve for Lymphocyte Volume SD.....	44
Figure 3.15: ROC Curves for all 6 Lymphocyte Statistical Parameters.....	45
Figure 3.16: Relation of Lymphocyte Volume SD to Manual Variant Lymphocyte Percent.....	46
Figure 3.17: ROC Curve of Lymphocyte Percent Parameter.....	47
Figure 3.18: Surface Rendering of LH750 WBC Differential Data of a Normal Donor.....	52
Figure 3.19: Eigensystem pose and variance descriptors for a single WBC population.....	57
Figure 3.20: Flow Chart Depicting Process of Extraction of Eigensystem-Based Descriptor Matrix.....	58
Figure 3.21: ROC Curve of Linear Classifier Using 7 Lymphocyte Statistical Parameters.....	61
Figure 3.22: Comparison of ROC Curves of Two Linear Classifiers.....	62
Figure 3.23: AUC Values for Classifiers Constructed from N-Tuples of Best Eigensystem Parameters.....	64
Figure 4.1: Conceptual Illustration of 3-Layer Perceptron ANN with Topology 8- 10-1.....	69
Figure 4.2: Accuracy, Sensitivity, and Specificity Results for 50 Training Trials – Experiment #1.....	75
Figure 4.3: Graph of Sigmoid Function Using Parameters Defined in Table 4.1.....	77
Figure 4.4: Accuracy, Sensitivity, and Specificity Results for 50 Training Trials – Experiment #2.....	78

Figure 4.5: Average Training Pattern Error versus Number of Hidden Units.....	80
Figure 4.6: Average Pattern Error versus Number of Hidden Units with Cross-Validation	84
Figure 4.7: Average Pattern Error versus Number of ANN Hidden Units.....	86
Figure 4.8: Accuracy, Sensitivity, and Specificity Results for 50 Training Trials – Experiment #5, Trial #1	87
Figure 4.9: Accuracy, Sensitivity, and Specificity Results for 50 Training Trials – Experiment #5, Trial #2.....	88
Figure 4.10: Accuracy, Sensitivity, and Specificity Results for 50 Training Trials – Experiment #5, Trial #3	89
Figure 5.1: Illustration of the cross-validation criterion to stop training.....	96
Figure 5.2: Variant Lymph Positive Sample – Discussion Pattern #1.....	98
Figure 5.3: Variant Lymph Negative Sample – Discussion Pattern #2	99
Figure 5.4: Variant Lymph Positive Sample – Discussion Pattern #3.....	101

SYMBOLS AND ABBREVIATIONS

AF	Activation Function
ANN	Artificial Neural Network
CCR	Correct Classification Rate
FN	False Negatives
FNR	False Negative rate
FP	False Positives
FPR	False Positive rate
MCR	Misclassification Rate
ROC	Receiver Operating Characteristics
AUC	Area under the Curve
PCA	Principal Component Analysis
STD	Standard Deviation
TN	True Negatives
TNF	True Negative fraction
TNR	True Negative rate
TP	True Positives
TPF	True Positive fraction
TPR	True positive rate
n1-n2-n3	ANN topology, where n1, n2, and n3 are the amount of neurons in input, hidden and output layers, respectively
L	Linear activation
S	Sigmoid activation

1. INTRODUCTION

Automated hematology analyzers are critical tools known for their ability to perform fast and accurate analyses of human blood. The aim of these analyses is to determine the possible existence of abnormalities in the human blood that have been correlated with serious disease states, such as infectious mononucleosis, leukemia, and various cancers. Though these analyzers do provide important feedback to medical experts, there is a persistent need to improve the sensitivity and specificity of the results, given the subtle nature of these diseases. This is evidenced by the occurrence in relatively high rates of both false positives and the more critical false negatives as reported by some users of these devices. Thus, it is advantageous to provide a superior pattern interpretation framework that is able to provide enhanced classification ability than is currently available.

For these reasons, the ultimate endeavor of this research was to improve the prospects for the automated classification, characterization, and diagnosis of white blood cells existing in various stages of maturation or exhibiting possible abnormalities. Toward this end, this dissertation proposes a research platform for the application of advanced feature extraction techniques to augment the accuracy of the classification and analysis of human blood cells that is performed by automated hematology analyzers such as the LH750TM that is currently manufactured by Beckman Coulter, Inc. Given the strong collaboration between FIU and Beckman Coulter, the research methodology and the practical implications of this dissertation focus on hematological research towards a better understanding of serious diseases such as leukemia and lymphoma.

In the context of automated hematological analysis, advanced feature extraction techniques and classification algorithms have been developed within this research platform to address the complex problems of detection and ultimately diagnosis of disease. Within the constraints of a real-time analysis system, achieving reliable accuracy for classification problems of this level of difficulty usually requires the use of high-performance computing systems that employ various levels of machine learning and artificial intelligence.

The term Artificial Intelligence is used to describe a collection of problem-solving methods that emerged after the Second World War. These methods were innovative because they used techniques to solve problems in engineering that were often unable to be solved using conventional mathematics [Fogel et al. 1966]. Advances in this field were, to a large degree, furthered firstly by the advent and later by improvements made in computing technology. As the use of computers became more commonplace, the use of AI principles in the areas of research, and problem solving has steadily grown. The areas of AI, which have made the most impact in the scientific community, are genetic algorithms, fuzzy logic, and artificial neural networks. Though the concepts of genetic algorithms and fuzzy logic are indisputably useful for problem solving, some have stated that neural networks have demonstrated an uncanny advantage due to their ability to learn.

The advent of the Perceptron was considered the beginning of the theory of Artificial Neural Networks (ANNs) [Rosenblatt 1958]. As the theory was further developed, other types of networks were also invented, such as the Adaline, Madaline, and Self-Organizing Maps [Specht 1990] [Widrow and Lehr 1990] [Kohonen 2001]. The most

attractive feature about ANNs is that they can mimic the way biological neurons interact to process and learn information. For these reasons, many problems of pattern recognition and classification can be solved with ANNs.

Even though AI concepts have been applied to many scientific and industrial fields, it is in the biomedical field where especially ANNs have clearly outperformed the other areas. In these areas, applications are mainly targeted to the recognition of patterns that are indicative of diseases, various forms of cancer, and tumors [Kothari et al 1996] [Zong et al 2010] [Reddick et al 1998].

Especially in hematological research, the main focus of this dissertation, the use of AI applications to deal with the automation of the detection of blood abnormalities from screening devices, such as hematology analyzers and flow cytometers, is becoming more prevalent. Early detection and diagnosis of diseases such as leukemia and lymphoma has become a critical area of research and has received great attention from medical and scientific institutions in the last two decades. This attention is due to the overwhelming number of persons suffering from leukemia, which reportedly affects over 300 thousand people in the United States, with an estimated 43,000 new cases diagnosed in 2010 alone. Likewise, lymphoma currently affects over 600 thousand people in the United States. The types of treatment and survival rates vary widely with the type of leukemia or lymphoma encountered, and can also differ by the person's age at diagnosis, gender or race [Leukemia and Lymphoma Society 2011].

This dissertation is devoted to the crucial role that advanced feature extraction and pattern recognition methods play in hematological research. For that reason, the strategies used to extract descriptive features from hematological blood data patterns

produced by the Beckman Coulter LH750 automated hematology analyzer will be presented. Additionally, this dissertation will present a classification framework used for detecting clinically abnormal levels of variant lymphocytes in human blood samples, attempting to improve the performance of this classification method by employing various design and optimization strategies, and ultimately showing the finalized design that demonstrated the best classification performance on the dataset used for this experiment.

Chapter 2 of this dissertation provides some fundamental background information about human blood cells, emphasizing the role that detection of white blood cell (WBC) abnormalities plays in diagnosing disease states. A basic description of the typical capabilities and usages of automated hematology analyzers in a clinical setting will be given. Following this will be a review of some existing methods of detecting hematological abnormalities in blood data, with an emphasis on the methods of extraction of descriptive features from blood data patterns that are currently implemented by the expert system software modules contained in currently-produced hematology analyzers. After a review of some existing detection methods, a discussion of the limitations of these methods will be described. This will provide a strong rationale for the necessity of the proposed multi-dimensional eigensystem-based feature extraction methodology.

Chapter 3 presents the experimental dataset that was acquired for the purposes of demonstrating that the correct classification of variant lymphocyte samples is an area of research that could benefit greatly from improved feature extraction methods. A survey of the basic commonly-used statistical features that are often used to develop variant lymphocyte classifiers is given. Special attention is also paid to the limitations of these

basic features; and it will be shown that, through the development of a pattern classifier that uses only the set of features derived from these basic statistics, that the two classes of samples (Variant Lymph Positive and Variant Lymph Negative) are not linearly separable, attesting to the difficulties that this problem inherently possesses. More importantly, these limitations show the immediate need for the development of a set of pattern descriptors that will provide an improved separation of the two classes, thus providing a potential solution with improved classification accuracy.

Chapter 4 will introduce the concept of multi-dimensional, eigensystem-based cluster descriptors and their potential applications to this classification problem. It will be demonstrated that the numerical descriptors extracted using this shape analysis paradigm represent a powerful set of features for the improved classification of hematological abnormalities. Naturally, following this introduction will be an in-depth description of the process that was designed to extract the eigensystem-based feature space from the raw hematological data files in a robust, yet computationally efficient way. It will then demonstrate, through the application of several machine learning implementations based on the Perceptron artificial neural network using the data provided for this study, that this set of descriptive features has the capability to classify samples with clinically relevant levels of variant lymphocytes more accurately than the existing classification methods used by Beckman Coulter on the LH750 automated hematology analyzer.

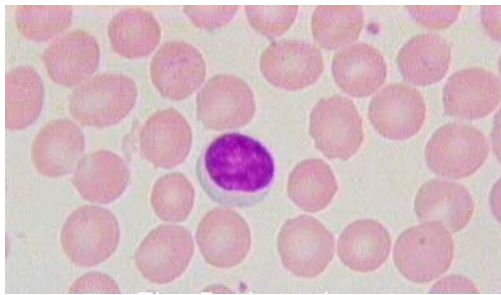
In Chapter 5, an in-depth discussion of the classification results will be made. In addition, extensions to the ideas proposed in this dissertation will be expressed, along some new potential shape descriptors that may have added utility for this classification problem. Concluding remarks are provided in Chapter 6.

2. BACKGROUND ON CLINICAL HEMATOLOGY AND BLOOD CELLS

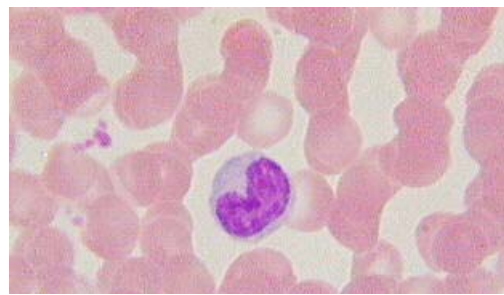
2.1. Background and Significance of Blood Cells

2.1.1. Basic Categories of Human Blood Cells

Hematology is defined as the study of blood. Blood consists of a liquid portion, called the serum or plasma, and its cellular components, the hematocytes. Mostly, the modern practice of hematology emphasizes the study of the hematocytes, or blood cells. Blood cells are generally divided into three main categories: Erythrocytes (Red Blood Cells), Leukocytes (White Blood Cells) and Thrombocytes (Platelets). For simplicity, the abbreviations WBC and RBC will be used to signify white blood cells and red blood cells, respectively. Though each of these cell types has its own significant functions in the human body, from the perspective of detection and diagnosis of many disease states, the white blood cells are often considered to be more indicative than the other two types. The white blood cells consist of five subcategories: Lymphocytes, Monocytes, Neutrophils, Eosinophils, and Basophils. These five cell types are illustrated in the following figures made from pictures taken from a microscopic view of the cells:



(a)



(b)

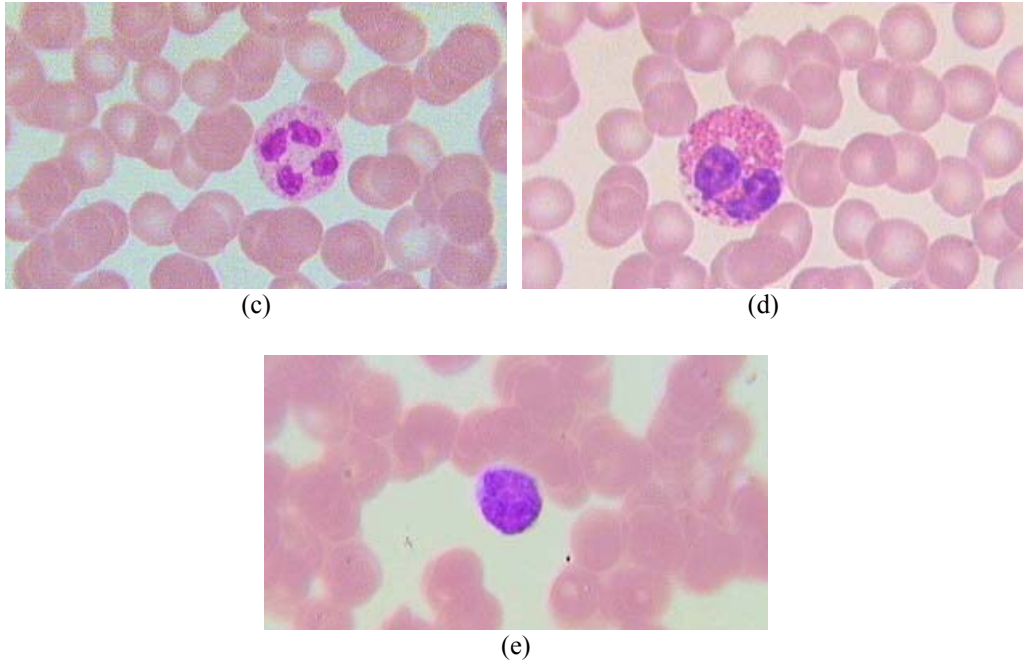


Figure 2.1: Basic WBC Types: (a) Lymphocyte, (b) Monocyte, (c) Neutrophil, (d) Eosinophil, (e) Basophil
 The reddish-to-purplish cell in each picture is the specific WBC cell type indicated, and the pink, semi-transparent cells in the background of each figure are Red Blood Cells (RBC).

2.1.2. Abnormal Variations in White Blood Cells

It is important to note that the types shown above represent normal, mature white blood cells, which are seen in blood under normal clinical conditions. However, in the event of disease or various illnesses, specific abnormal white blood cells may also appear. In this context, the term “abnormal” could refer to a white blood cell that is not fully mature (immature) or has physical attributes or characteristics that are noticeably different than its normal, mature counterpart [Bessman 1986]. Some examples of abnormal/immature WBCs are the following: Variant/Atypical Lymphocytes, Blasts, Immature Bands, Immature Granulocytes, and Plasma Cells.

Each of these abnormal/immature cell types is briefly defined, as follows:

- Variant or Atypical Lymphocytes: Lymphocytes showing morphologic features different from normal lymphocytes. These lymphocytes are often larger than normal lymphocytes due to antigen stimulation. In addition, their nuclei can be round, elliptic, indented, cleft, or folded. Their cytoplasm is often abundant and can be basophilic with vacuoles and/or azurophilic granules present.
- Blasts: The most immature, undifferentiated cells in the WBC lineage. These cells are usually devoid of definitive characteristics that would identify them as belonging to a particular WBC cell sub-type.
- Immature Granulocytes: An immature neutrophil that may be neutrophilic, acidophilic, or basophilic in character.
- Immature Bands: An immature neutrophil that has a band-shaped or horseshoe-shaped nucleus.
- Plasma Cells: Lymphocytes that were produced in the bone marrow that are specialized for antibody (immunoglobulin) production.
- It is due to the detection of cellular variations, or deviations from the normal, that allows clinical hematology the ability to be useful as a diagnosis method.

2.1.3. Benefits of the WBC Manual Differential

Historically, the observation of WBCs and distinctions between normal and abnormal WBCs has been routinely made by trained hematologists using microscopic visual inspection techniques. This process of microscopic examination of the white blood cells (and their separation into respective categories) by a trained hematologist or doctor is known as the *Manual WBC Differential* and is currently still considered an invaluable

source of hematological information. The usual process of a manual differential is to examine a fixed number of WBCs (usually a tractable number in the range of 100-400 cells) and then tabulate the number of cells of each WBC type that were encountered, as exemplified in Table 2.1 for a 200-cell WBC manual differential analysis.

Table 2.1: Example of a 200-Cell Manual WBC Differential Count Report

Lymph	Mono	Neutro	Eo	Baso	Blast	Immature Gran	Immature Band	Variant Lymph	Plasma Cell
105	3	84	1	1	0	0	2	4	0

Naturally, these values are often normalized to their respective percentages using the total number of WBCs counted (in this case 200 cells), to yield the percentages in Table 2.2

Table 2.2: Example of a 200-Cell Manual WBC Differential Percentage Report

Lymph	Mono	Neutro	Eo	Baso	Blast	Immature Gran	Immature Band	Variant Lymph	Plasma Cell
52.5%	1.5%	42%	0.5%	0.5%	0%	0%	1.0%	2.0%	0%

Of particular interest is the distribution of percentages across the various WBC sub-categories represented. The reason is that each of the many sub-populations has a different level of medical significance. Even though microscopic visual inspection of cells is still considered the most reliable “Golden Standard” of identification, this method is not without its limitations [Bessman 1986] [Drouet and Lees 1993].

2.1.4. Limitations of the WBC Manual Differential Process

Firstly, it is very time consuming to perform manual cell differentials. Often a single manual reader will only be able to view and classify perhaps 200 cells in a single sitting before visual and physical fatigue may force the reader to need a rest. For this reason, a single manual reader can usually only perform at most 25 manual differential procedures in a typical work day. In fact, to assure quality and accuracy of the readings made for manual differentials; many medical institutions have recommended limits on their personnel as to the maximum number of manual differentials they should attempt in a single work day [Bessman 1986]. From the standpoint of monetary expenditures due to labor costs of employing highly-trained individuals such as cellular biologists and hematologists, one can easily deduce that the manual differential procedure can be rather costly.

Of paramount importance in the WBC differential are not only the distributions of percentages of normal, mature WBC sub-types (Lymph, Mono, Neutro, Eo, and Baso), but also the percentages of abnormal/immature types (Blasts, Immature Granulocytes, Immature Bands, Variant Lymphocytes, and Plasma Cells). Since there is a multitude of known types and variations of WBC cells (both normal and abnormal), correct and consistent identification of all the cells seen on a patient's blood smear can be rather difficult, even for technologists and medical doctors with years of training and experience. In fact, since some of the cell types encountered share many common morphological attributes, shapes, and colors, it is often very subjective as to the correct classification of some cells. Due to problems such as time and cost constraints of performing manual differentials, and the propensity for human subjectivity in blood cell

classification, automated hematology analysis devices were created to overcome some of these shortcomings.

2.2. Automated Hematology

2.2.1. Automated Hematology Devices

Up until the middle of the 20th century, the blood smear slide reading with an optical microscope had been the only clinical method for blood analysis. However, this began to change in the early 1950s. A landmark event happened in 1953, when Wallace Coulter invented the “Coulter Principle”. This principle basically demonstrated that a “particle pulled through an orifice, concurrent with an electric current, will produce a change in impedance that is proportional to the volume of the particle traversing the orifice” [Rodak et al 2007]. This methodology was eventually applied to count and measure the volume of individual blood cells suspended in a saline solution, and the first Coulter blood cell counter was subsequently born, thus making the dream of automated blood analysis into a reality.

The advent of computers in the 1960’s helped to bring automated blood analysis into a new era. Thus, signal processing, representation and statistical classification methods could then be applied to the acquired blood cell measurements, while seeking more accurate measurements. Concurrently, developments made in the physics, clinical chemistry and immunology fields led to the advent of using lasers in hematology analyzers so that additional cellular measurements such as light scatter and fluorescence could be acquired, which were demonstrated to add even more selectivity to automated blood cell classification methods.

From the 1990's to the present, automated blood analyzers have adopted technologies that employ fluorescent antibodies, which can be developed to be chemically selective to specific cell types. Thus, current analyzers can acquire a set of parameters to represent different aspects of each cell, such as volume, conductivity, light scatter, and multiple fluorescence wavelengths. Even though technological advances have been made recently, presently over 98% of commercially produced automated blood cell analyzers still incorporate the basic elements of the Coulter principle into their designs [Rodak et al 2007].

2.2.2. VCS Automated Hematology Devices

Within the experimental set up of this dissertation, the research carried out using data collected from the Beckman Coulter LH750 Hematology Analyzer. This analyzer is known as a "VCS" analyzer, where the letters V, C, and S are abbreviations for Volume, Conductivity, and Scatter, respectively. The formal descriptions of these parameters and their uses in cellular measurement are defined as follows:

V: A measurement proportional to the physical volume of the blood cell.

C: A measurement proportional to the electrical conductivity of the blood cell.

S: A measurement proportional to the light scatter produced by the blood cell.

Since the white blood cells are usually more indicative of potential disease states, interpretation and analysis of data produced by the Beckman Coulter VCS White Cell differential module (known as DIFF) is the sole focus of this research endeavor.

2.2.3. Uses of Automated Hematology Analyzers As Screening Tools

There is a multitude of disorders that can cause abnormalities in the White Blood Cells, some of the more common disorders are: sepsis, infectious mononucleosis, lymphomas,

dysplastic syndromes, viral infections, parasitic disorders, and various types of leukemia [Turgeon 2005]. In fact, a breakdown of the most commonly encountered hematological malignancies seen in the United States [Horner et al 2009] is given in Table 2.3.

Table 2.3: Breakdown of Hematological Malignancies in the United States

Type of Hematological Malignancy	Percentage
Acute Lymphoblastic Leukemia (ALL)	4.0%
Acute Myelogenous Leukemia (AML)	8.7%
Chronic Lymphocytic Leukemia (CLL)	10.2%
Chronic Myelogenous Leukemia (CML)	3.7%
Acute Monocytic Leukemia (AMOL)	0.7%
Other Leukemias	3.1%
Hodgkin's Lymphomas (all four subtypes)	7.0%
Non-Hodgkin's Lymphomas (all subtypes)	48.6%
Myelomas (all types)	14.0%
Total	100%

Depending on the type of disease and its severity, its level of progression, and whether or not treatments (chemotherapy, bone marrow transplants, or medications) have been administered to the patient, the hematological pattern can vary greatly. In fact, hematological information from automated analyses used alone is not recognized by medical doctors as sufficient information for medical diagnosis. It is merely used as a single source of information in the process of differential diagnosis, which often employs

the use of a multitude of other diagnostic tests and analyses [Turgeon 2005] [Drouet and Lees 1993].

2.2.4. Examples of Common Hematological Patterns

In the context of the Beckman Coulter LH750 automated hematology analyzer, it is often visually apparent through the displayed hemograms, when a disorder or abnormality exists within a patient's blood. The hematological patterns for a clinically normal blood donor run with the LH750 instrument in the DIFF mode (Automated White Cell Differential) are shown in Figure 2.2.

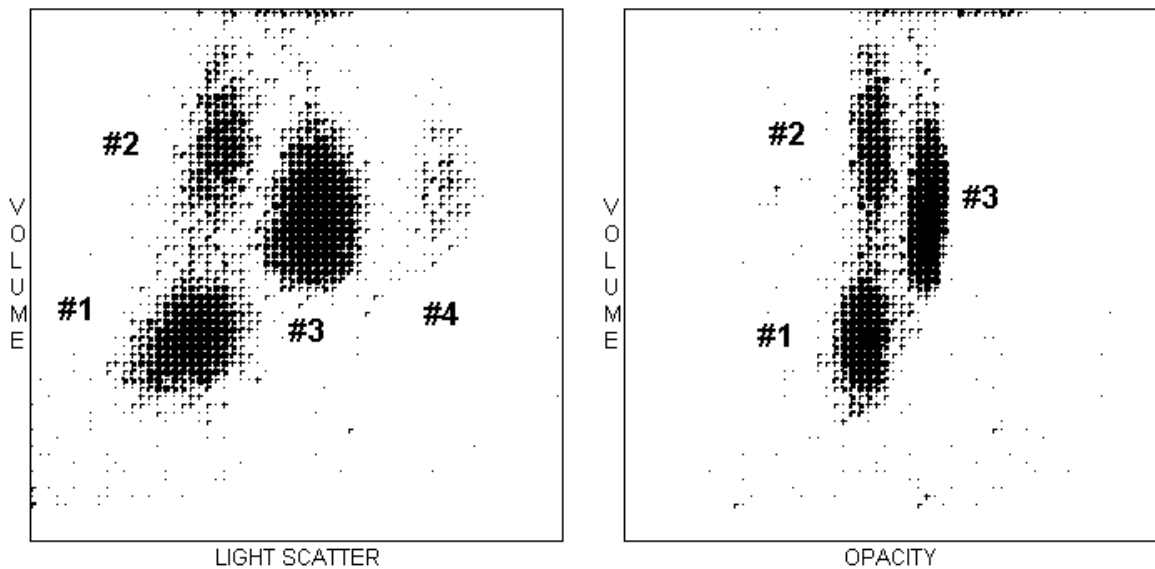


Figure 2.2: VCS WBC Differential Pattern of Normal Blood Donor – Before Event Categorization

The two views shown (Volume versus Light Scatter and Volume versus Opacity) are two-dimensional histogram representations of the measurement parameters collected for the patient's blood cells that were analyzed by the LH750 instrument. Note that the numbers 1-4 have been drawn on top of these plots; this was done for the purposes of discussion. On the left 2D scatterplot, there are 4 distinct clusters or populations present, each with the numbers #1-4 next to them. These are the Lymphocyte, Monocyte,

Neutrophil, and Eosinophil populations, respectively. From earlier discussions in this section, it is known that these are separate types of mature WBC cells. In the right 2D scatterplot, it should be noted that only 3 distinct clusters are present; this is because the Neutrophil and Eosinophil populations overlap completely in the Opacity dimension. Thus, on the right 2D scatterplot, population #1 is Lymphocytes, #2 is Monocytes, and #3 is Neutrophils and Eosinophils overlapped.

The automated software module in the analyzer then separates the data points into their respective categories using a complex, statistical, rule-based clustering and segmentation scheme. An example of the segmented patterns into their mutually-exclusive categories is as shown in Figure 2.3.

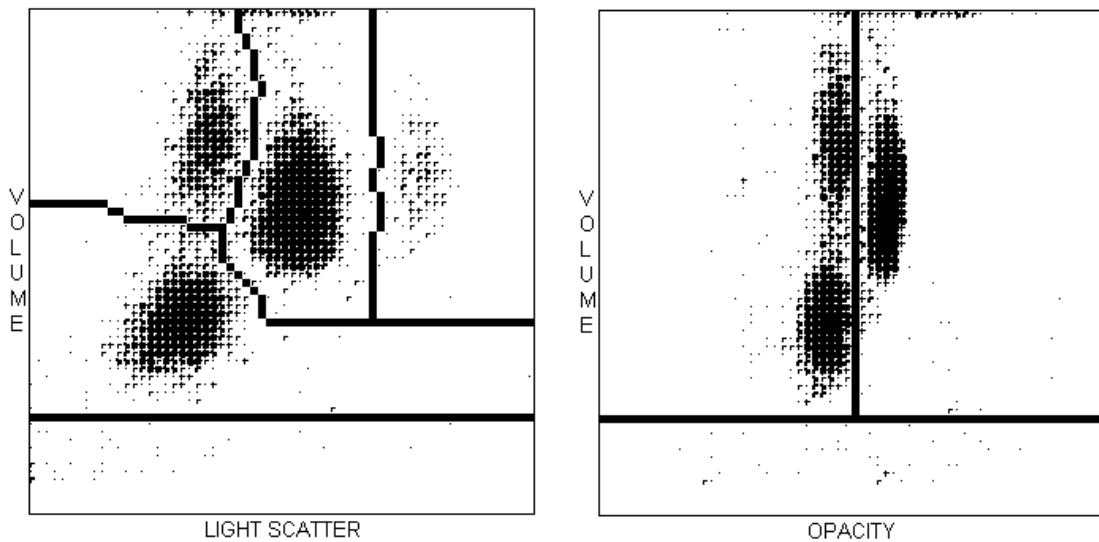


Figure 2.3: VCS WBC Differential Pattern of Normal Blood Donor – After Event Categorization

Also worth mentioning are the events that are below the segmentation line beneath the lymphocyte population; these are non-WBC events such as residual red blood cells and platelets. These events are usually termed debris. Additionally, no mention of the basophil population was yet mentioned. This is because the basophil events are normally

very rare, and do not often form a distinct cluster of events. However, when a significant amount of basophils is present, its events can be seen as a cluster to the right of the segmentation line that is to the right of the lymphocyte population on the Volume versus Opacity axis.

There are many distinct patterns that can be seen, caused by a variety of disorders and diseases. Some of these diseases and their respective hematology scattergram patterns, as depicted by the LH750 hematology analyzer's WBC differential, or "DIFF" software module, are the following:

Infectious Mononucleosis (IM):

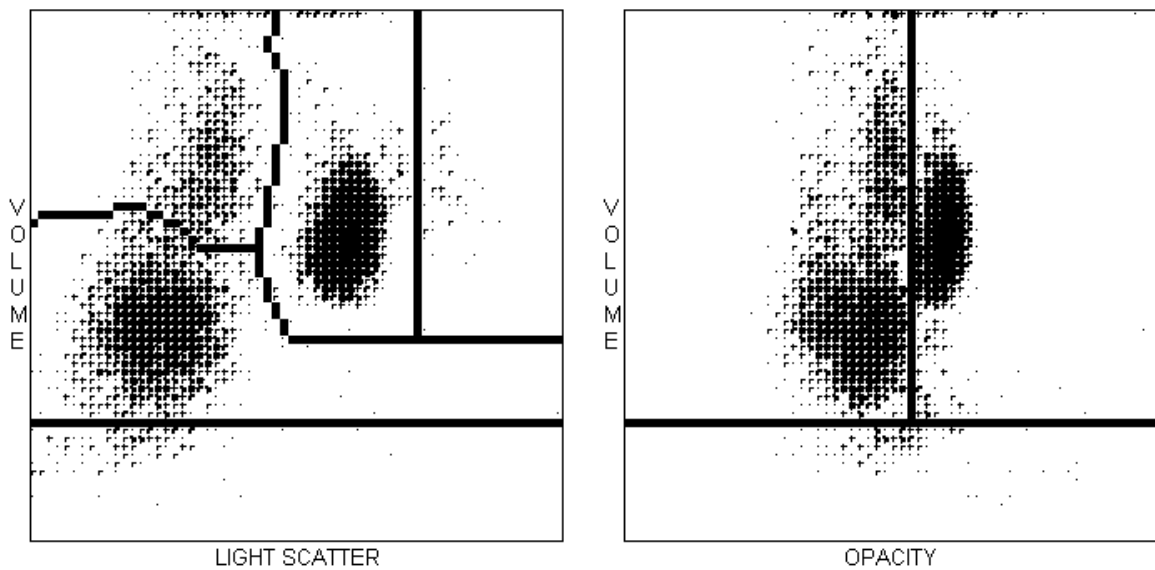


Figure 2.4: WBC Differential Patterns of Donor Diagnosed with Infectious Mononucleosis

Infectious Mononucleosis is one of the more common disorders that cause an abnormal increase in the level of variant lymphocytes in the blood. Even though this disorder can have dangerous consequences, it is usually regarded by medical experts as being on a lower level of criticality than malignant disorders such as leukemias and lymphomas [Caldwell and Lacombe 2000].

Chronic Lymphocytic Leukemia (CLL):

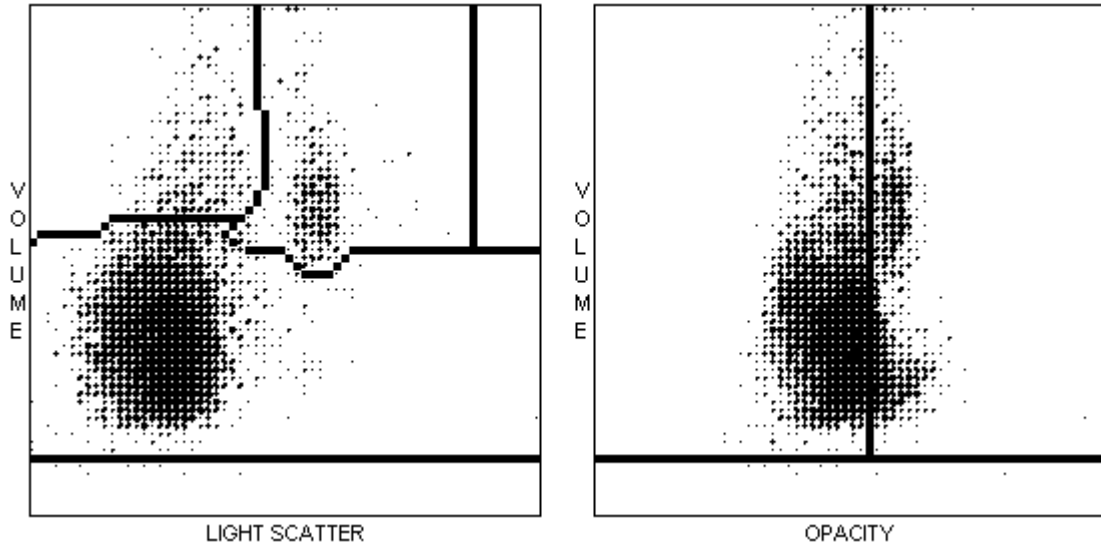


Figure 2.5: WBC Differential Patterns of Donor Diagnosed with Chronic Lymphocytic Leukemia

Chronic Lymphocytic Leukemia is a disorder that is quite prevalent. Normally, this disorder is more common in male patients over 50 years of age; children almost never develop CLL [Caldwell and Lacombe 2000]. An important fact is that CLL samples are sometimes difficult to distinguish from samples with abnormal levels of variant lymphocytes because both of these types of samples share some common characteristics. Both types of patterns usually exhibit lymphocytosis, and, at the same time, large or elongated lymphocyte populations – thus making them both have abnormal population statistics.

Acute Lymphocytic Leukemia (ALL):

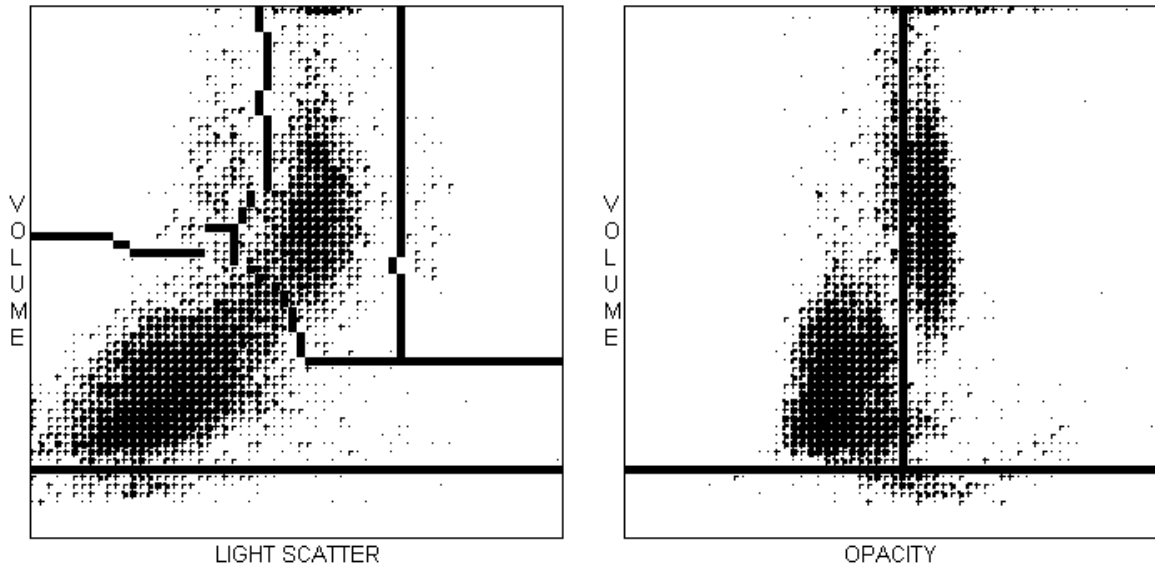


Figure 2.6: WBC Differential Patterns of Donor Diagnosed with Acute Lymphocytic Leukemia

This type of sample also exhibits abnormal lymphocyte statistics (elongated lymph population and often lymphocytosis), thus also making it a difficult pattern to differentiate from a variant lymphocyte sample. This type of sample usually contains very immature lymphocytes, also called lymphoblasts. Flagging this type of sample as containing variant lymphocytes, which would technically incur a false positive, is usually an acceptable error, since it is indeed an abnormal pattern that should be detected.

Acute Myelocytic Leukemia (AML)

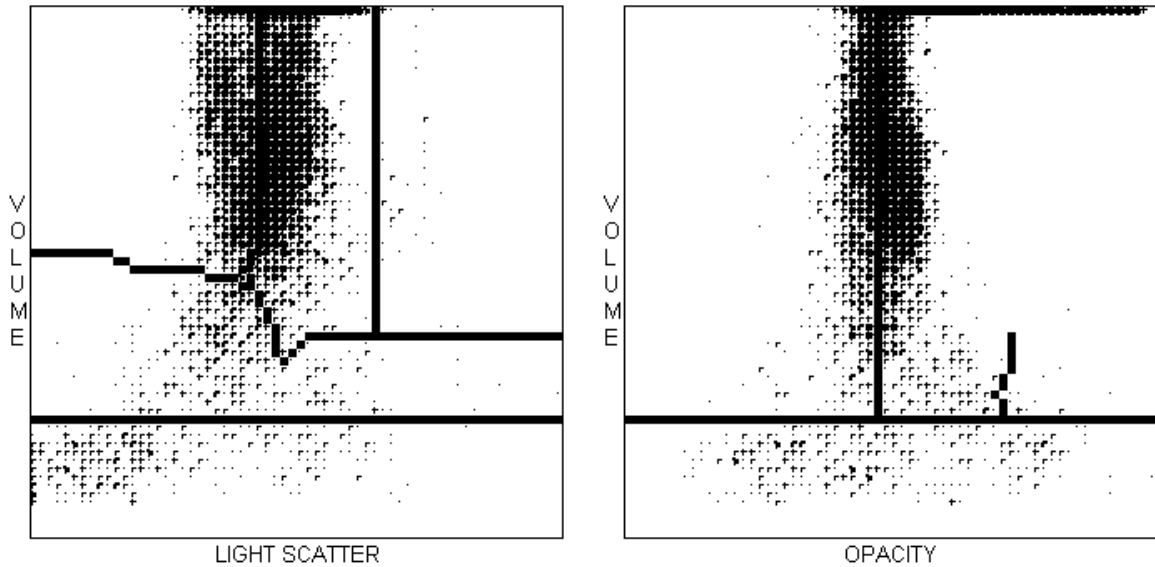


Figure 2.7: WBC Patterns of Donor Diagnosed with Acute Myelocytic Leukemia

A sample such as AML (Acute Myelocytic Leukemia) usually has a hematological pattern that is abnormal; the abnormalities in these types of cases affect mostly the monocyte and neutrophil populations, the lymphocyte population is normally decreased in size or is virtually non-existent in such cases. However, there are many varieties of myelocytic disorders that exist and a thorough discussion of all of them would be beyond the scope of this dissertation.

2.2.5. Limitations of Automated Hematology Analyzers

Unfortunately, all abnormal patterns do not express such a visually obvious departure from that of a normal blood pattern, as the ones shown in the previous section. Some patterns may look normal, yet may have abnormalities or malignancies present, thus elucidating the subtle nature of the problem at hand. Usually, the greater the proportion of abnormal or malignant cells that are present in a population or region of the pattern, the more the population or region begins to change in appearance. One type of abnormal

cell that has traditionally been very hard to detect reliably is the Variant or Atypical Lymphocyte, especially when the percentage of Variant Lymphocytes, as a fraction of total WBC cells, is less than 10% [Aulesa et al 2003] [Aulesa et al 2004]. Variant lymphocytes can be found in increased numbers (usually proportions greater than 10% but sometimes much more) in disorders such as infectious mononucleosis, viral pneumonia, and viral hepatitis. However, even clinically normal donors can have up to 4% variant lymphocytes present, thus making the delineation between clinically normal and clinically abnormal more difficult to create [Turgeon 2005] [Van der Meer et al 2007] [Caldwell and Lacombe 2000].

In fact, several publications have been written on the relatively low accuracy of Variant Lymphocyte detection using several different manufactured automated hematology analyzers [Hoffmann and Hoedemakers 2004]. In particular, the variant lymph flag on the Coulter LH750 automated hematology analyzer, for instance, has been evaluated by some external studies and shown to have limited accuracy, specificity, and sensitivity for consistently detecting medically relevant levels of variant lymphocytes in whole blood samples [Aulesa et al 2004].

For this reason, this dissertation will focus on improving the detection of variant lymphocytes, since this type of pattern classification currently demonstrates some limitations and difficulties.

2.3. Overview of Methods for Detection of Abnormal Hematological Patterns

2.3.1. Basic Operational Steps of Existing Classification Methods

The simplest detection of hematological pattern abnormalities is performed using the statistical parameters (percents, means, and standard deviations) derived from the events

in each population followed by rule-based classifiers applied on these parameters. All detection methods have in common the need to “learn” the conditions under which a pattern may be considered abnormal. Rules of this type are often developed using sets of data that are representative of normal donors and abnormal donors; then average statistics for each population are created. Then, the differences in each statistic between the two datasets would be examined and the subsequent statistics offering the largest separation (the largest difference between normal and abnormal donors) would be kept as candidate features for rule development. After the rule base has been extracted with some method, the algorithm is tested on unknown data to evaluate its performance.

The steps undertaken in the context of this dissertation can be generalized in the following way:

- 1) Data collection and pre-processing (including data partitioning for pattern extraction and later testing)
- 2) Feature extraction
- 3) Rule extraction (training)
- 4) Event detection/classification (testing)

There are many ways of detecting such patterns: either by rule extraction or by using more sophisticated methods such as artificial neural networks (ANNs).

The patterns of interest to a particular research study often can vary greatly, and thus are usually extracted using a specific technique that has been customized for the task at hand. Once the set of useful features has been acquired, these values are then incorporated into a classification framework. For example, an implementation of a perceptron ANN to classify hematological data as normal or abnormal, using an 84-dimensional set of

features produced by the Advia 120 automated hematology analyzer, was done recently by a group of researchers, with a claimed efficiency of classification of 91% [Zini and D’Onofrio 2003].

When comparing detection or classification algorithms in terms of performance, some sort of criterion is needed in order to select the best algorithm. Testing such algorithms produces different types of errors, which can make comparisons difficult. Thus, the next section will describe methods to evaluate algorithms in terms of performance.

2.3.2. Method of Performance Evaluation of Detection Algorithms

Evaluation of the performance of detection algorithms is usually performed using Receiver Operating Characteristics (ROC) Analysis [Tilbury et al 2000]. Such analysis begins by establishing a confusion matrix which contains information about the actual classification of the data being tested and the outcome of the classification system. Table 2.4 shows the main entries of the confusion matrix for a two-class classifier.

Table 2.4: Entries of a confusion matrix

		Detected as	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

The four table entries are defined as follows: TP (true positives) is the number of correct classifications that an instance is positive; FN (false negatives) is the number of incorrect classifications that an instance is negative; FP (false positives) is the number of incorrect classifications that an instance is positive; and TN (true negatives) is the number of correct classifications that an instance is negative.

Positive and negative refers to the outcome given by the classifier, whereas true and false refers to the correctness of this outcome (i.e. right or wrong with respect to the actual state of the patient). The sum of the first and second row is the total number of positive and negative instances being under test, respectively, whereas instances are just all data values to be classified, regardless of their class. Similarly, the sum of the first and second column is the total number of positive and negative detections by the system, respectively. The grand total is the total number of instances being classified.

Rows and columns summarize to the following:

- N_{neg} is the total number of negative instances: $N_{neg} = TN + FP$ (2.1)

- N_{pos} is the total number of positive instances: $N_{pos} = TP + FN$ (2.2)

- C_{neg} is the total number of negative classifications: $C_{neg} = TN + FN$ (2.3)

- C_{pos} is the total number of positive classifications: $C_{pos} = TP + FP$ (2.4)

- N_{tot} is the total number of instances being detected:

$$N_{tot} = N_{neg} + N_{pos} = C_{neg} + C_{pos} = TP + TN + FP + FN \quad (2.5)$$

The following quantities can be extracted from the confusion matrixes:

- Correct Classification Rate (CCR, also called accuracy): The proportion of all correct classifications to the total number of instances:

$$CCR = \frac{TP + TN}{N_{tot}} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.6)$$

- Misclassification Rate (MCR): The proportion of all incorrect classifications to the total number of instances:

$$MCR = \frac{FP + FN}{N_{tot}} = \frac{FP + FN}{TP + TN + FP + FN} = 1 - CCR \quad (2.7)$$

- True Positive Fraction (also known as Sensitivity): The proportion of TP to the total number of positive instances:

$$TP_f = \frac{TP}{N_{pos}} = \frac{TP}{TP + FN} \quad (2.8)$$

- True Negative Fraction (also known as Specificity): The proportion of TN to the total number of negative instances:

$$TN_f = \frac{TN}{N_{neg}} = \frac{TN}{TN + FP} \quad (2.9)$$

- True Positive Rate (also known as Precision): The proportion of TP to the number of positive classifications:

$$TP_r = \frac{TP}{C_{pos}} = \frac{TP}{TP + FP} \quad (2.10)$$

- True Negative Rate: The proportion of TN to the number of negative classifications:

$$TN_r = \frac{TN}{C_{neg}} = \frac{TN}{TN + FN} \quad (2.11)$$

- False Negative Rate: The proportion of FN to the number of positive instances:

$$FN_r = \frac{FN}{N_{pos}} = \frac{FN}{TP + FN} = 1 - TP_f \quad (2.12)$$

- False Positive Rate (False Alarm Rate): The proportion of FP to the number of positive instances:

$$FPR = \frac{FP}{N_{neg}} = \frac{FP}{TN + FP} \quad (2.13)$$

Concerning the values of FNR and FPR, low values are usually desired. When two classifiers are being compared, accuracy is often considered the most important metric, however, especially in the medical field, it is also important to monitor the rates of TP and TN as well. For these cases, measures such as sensitivity or specificity should be maximized in addition to accuracy.

Sometimes, it is desired to tune a classifier's performance to meet required performance metrics, but it is known that varying a classifier's threshold, or operating point, can have contradictory effects. For instance, altering a classifier's operating point to increase the TP rate can have the undesired effect of increasing the FP rate. This variation of classifier performance as this threshold is varied is best given by a Receiver Operating Characteristic (ROC) curve [Marcum 1960]. The ROC curve is a parametric curve that is constructed based on the values of the TP and FP rate. In Figure 2.8, two ROC curves from two different classifiers are plotted. Observe that the classifiers excel each other in different regions of the plot.

Depending on the problem at hand, any particular ROC measure can be chosen in favor of another. For example, if one prefers to maximize TP and minimize FN, then TPF is the best choice. Often, a compromise is made by using the area under the ROC curve, and the classifier with the highest area under the ROC curve is said to be the best.

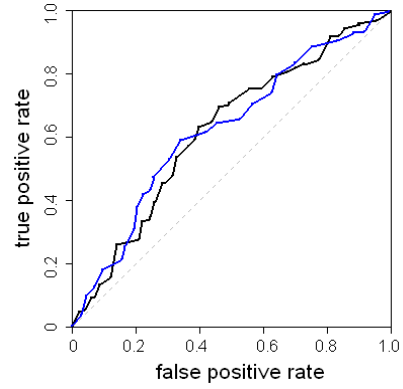


Figure 2.8: Two ROC curves plotted from the confusion matrixes of two classifiers

This usage of ROC curves will be discussed in Chapter 3, where it will be used for numerical evaluation of potential features for their abilities to effectively separate data into classes.

3. FEATURE EXTRACTION METHODS FOR VARIANT LYMPHOCYTE CLASSIFICATION

3.1. Problem Introduction

Since the detection of variant lymphocytes was indicated as an area of difficulty for the classification and analysis rule-engines of many automated hematological analyzers produced by such reputable manufacturers such as Advia, Beckman Coulter, and Sysmex [Aulesa et al 2004] [Hoffmann and Hoedemakers 2004], it seemed obvious that some of the shortcomings in detecting this type of hematological disorder might be alleviated by an improved feature extraction framework.

In most hospitals where routine pre-screening of patients is often done, automated hematology analyzers are relied upon for their ability to detect the types of abnormal patterns discussed in the previous section. For patients with diseases that are in the acute stages, or have progressed to later stages where the blood or bone marrow are pushing out immature cells at a very high rate, often the patterns express a very obvious abnormal appearance, and usually, a measure of population width or elongation, such as the standard of deviation, can be used in isolation to flag the pattern.

This is evident from Figures 3.1 and 3.2, one representing a normal donor, and the other a donor with lymphocytic leukemia:

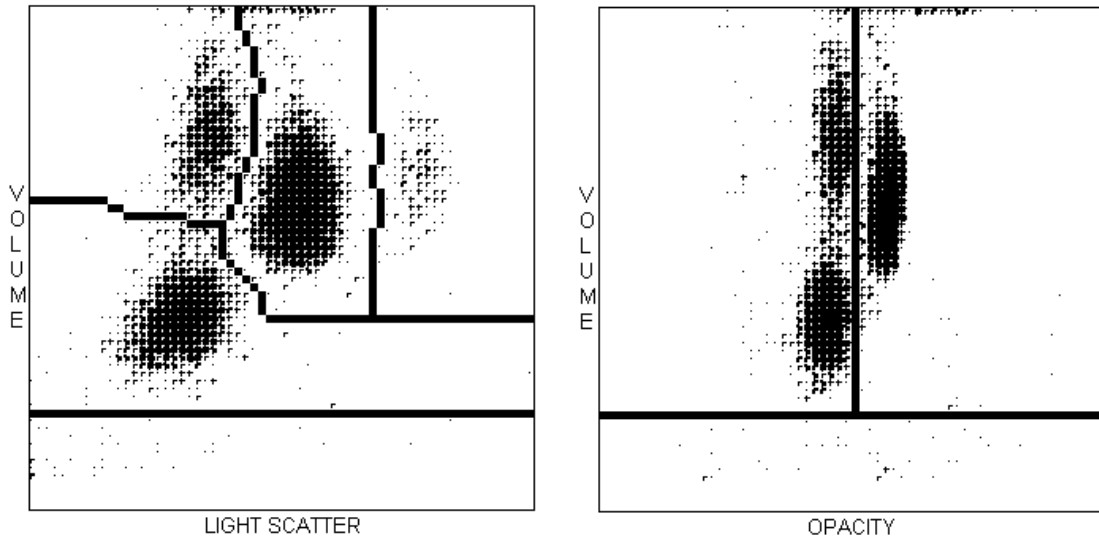


Figure 3.1 VCS WBC Differential Pattern of Normal Blood Donor

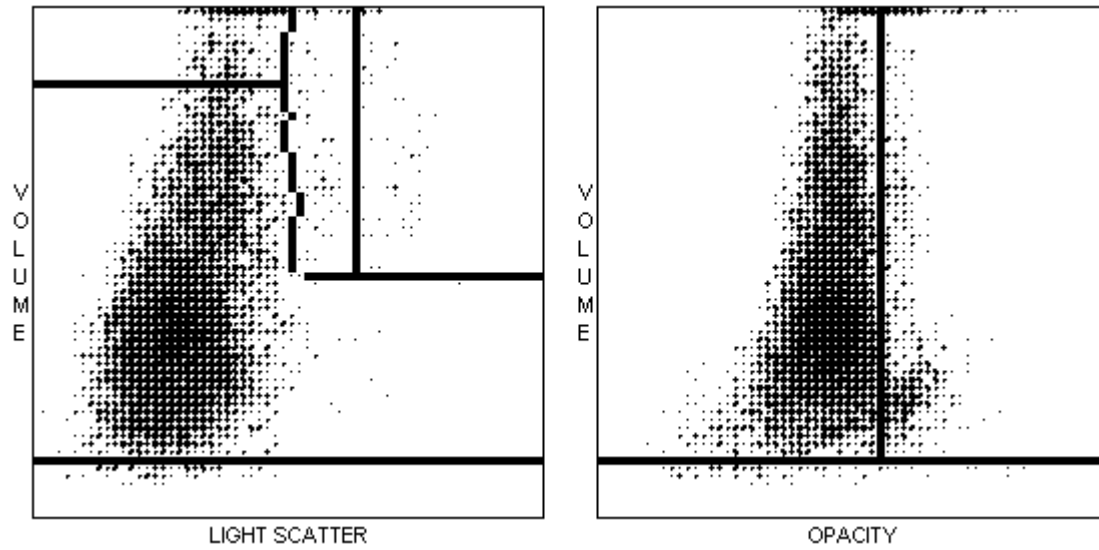


Figure 3.2 VCS WBC Differential Blood Pattern of Blood Donor with Lymphocytic Leukemia

A striking difference exists between the appearances of these two patterns. One can observe that the normal donor has 4 distinct populations present, which are the Lymphocytes, Monocytes, Neutrophils, and Eosinophils. In contrast, Figure 3.2 shows a

dominant lymphocyte population , with only small numbers of remaining events in the other three population categories.. This basic hematological condition is known as lymphocytosis (an over-abundance of lymphocytes). This, in fact, is usually the most common medically-relevant trait that is seen in cases of potential lympho-proliferative (diseases affecting the lymphocyte population) disorders, such as lymphocytic leukemia [Caldwell and Lacombe 2000]. Of strong clinical relevance also is the striking size of the lymphocyte population, and its expanse (or variance) in the Volume dimension. In fact, for this sample, this is an obvious feature detection trait. This difference is demonstrated in Figure 3.3:

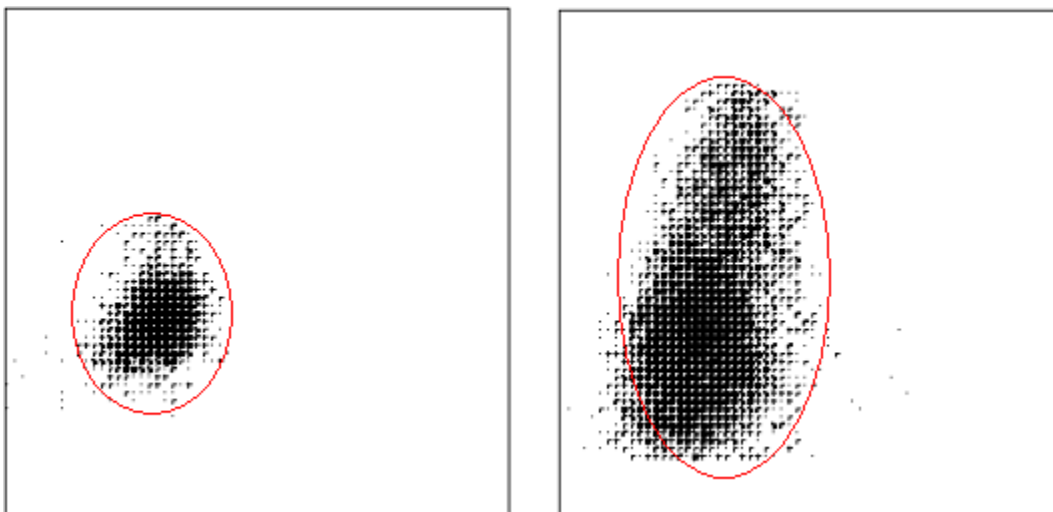


Figure 3.3: Normal (left) and Lymphocytic Leukemia (right) Lymphocyte Populations

The ellipses drawn around each lymphocyte population are used to indicate the variance or spread of each population in the Volume dimension; it is apparent that the lymphocyte population on the right (the Lymphocytic Leukemia sample) has a volume variance that is roughly three times as large as the lymphocyte volume variance of the lymphocyte population on the left (the Normal donor). Of course, not all abnormal or leukemic

patterns exhibit this behavior; hence a measure of caution is required when classifying these types of patterns.

In the immediate sections that follow, additional information is provided about the types of patterns that are indicative of abnormal levels of variant lymphocytes, and especially, about the features that are often used to detect them.

3.2. Variant Lymphocyte Experimental Dataset

Since the primary aim of this dissertation is to perform a feature analysis and development of an improved variant lymphocyte detection method, a specific dataset that is representative of this disorder was chosen for this study. The dataset consisted of 300 individual donors with the following breakdown:

1. 150 donors whose blood contains a clinically abnormal level of variant lymphocytes.
2. 150 donors whose blood contains a clinically normal level of variant lymphocytes.

All 300 raw data files used for this study were collected by trained technicians from a Beckman Coulter LH750TM automated hematology analyzer. To establish whether a “clinically normal” or “clinically abnormal” level of variant lymphocytes is present in a donor’s blood, a microscopic 200-cell WBC manual differential count performed by a qualified medical professional was provided for each donor to establish the true percentage of variant lymphocytes in each donor’s blood. This percentage of variant lymphocytes is considered the “Golden Standard” reference method, against which the accuracy of all proposed classification methods will be gauged.

In several published medical texts, there is an ongoing controversy about the appropriate significance level at which the proportion of variant lymphocytes should be considered medically abnormal in any given patient. Older medical standards such as the NCCLS (National Committee for Clinical Laboratory Standards) had previously set the accepted variant lymphocyte level at 10% (thus, a variant lymphocyte proportion of 10% or greater seen in a patient's WBC differential count should be considered medically abnormal). However, more recent medical standards established in 2007, such as the CLSI (Clinical and Laboratory Standards Institute) have revised this threshold, declaring that a level greater than or equal to 5% variant lymphocytes should be considered medically relevant [Koepke et al 2007]. Therefore, in accordance with the latest accepted medical standard, as established by the CLSI, a level of 5% or greater will be considered in this dissertation the reference positive threshold for variant lymphocytes.

3.3. Case Studies of Typical Variant Lymphocyte Hematological Patterns

In order to appreciate the intricacies involved in the subtle tasks of feature extraction and pattern classification, some commonly encountered patterns containing significant levels of variant lymphocytes are given as initial indicators of the difficulties that are yet to be resolved but need to be overcome.

Some disease states, especially leukemias and certain lymphomas, often have fairly recognizable WBC scattergram patterns. For this reason, reliable pattern analysis and recognition frameworks can be useful for their automatic detection [Caldwell and Lacombe 2000]. With a specific focus on disorders involving the lymphocyte population,

some of the more common diseases that can exhibit abnormal levels of variant lymphocytes and usually have visually-apparent abnormal hematological scattergram patterns are shown in the following case studies:

1. Case #1: Medically-Positive Variant Lymphocyte Sample

The following case represents a sample with a medically-positive level of variant lymphocytes which ideally, should be flagged as containing Variant Lymphocytes. This pattern is shown below:

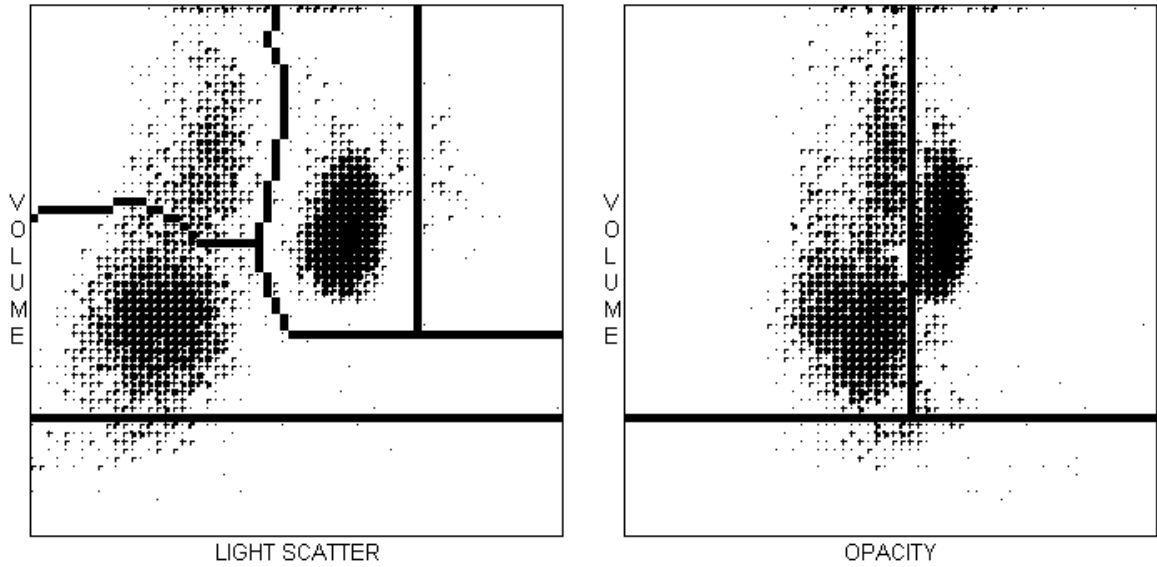


Figure 3.4: Blood Pattern with Abnormal Level of Variant Lymphocytes – Case #1

The percentages of the WBC population reported for this sample by the LH750 differential algorithm are given in Table 3.1.

Table 3. 1: LH750 Differential Algorithm Reported WBC Percentages for Case #1

Lymph%	Mono%	Neutro%	Eo%	Baso%
34.06%	9.72%	55.33%	0.68%	0.214%

Likewise, the statistical parameters (volume mean, volume standard deviation, opacity mean, opacity standard deviation, light scatter mean, and light scatter standard deviation) for the most pertinent WBC populations are shown in Figure 3.5.

	NE		LY		MO		EO	
	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
V	149.88	18.98	104.37	20.42	187.85	26.92	171.06	22.07
C	152.41	6.32	113.49	15.55	123.97	7.32	156.22	13.97
S	150.70	10.62	61.37	17.02	86.68	10.84	199.56	8.35

Figure 3.5: VCS WBC Population Statistical Parameters – Case #1

Likewise, the Manual WBC differential percentages for this donor are given in Table 3.2.

Table 3.2: Manual Differential WBC Percentages for Case #1

Lymph	Mono	Neutro	Eo	Baso	Blast	Immature Gran	Immature Band	Variant Lymph	Plasma Cell
23.0%	6%	59%	0.5%	0.5%	0%	0%	1%	10%	0%

Since this donor has a medically relevant abnormality (variant lymphocytes = 10% as a fraction of total WBCs), presumably a difference between the population statistics of this donor and a normal donor should be fairly evident. The VCS WBC population statistics from the normal blood sample are shown in Figure 3.6

	NE		LY		MO		EO	
	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
V	158.63	21.68	98.75	14.72	191.62	21.10	169.53	19.48
C	144.72	6.17	114.55	6.91	119.33	4.56	144.13	7.92
S	135.77	10.30	76.45	12.59	90.69	8.24	197.21	8.10

Figure 3.6: VCS WBC Population Statistical Parameters - Normal Blood Sample

If a percentage difference is calculated for each statistical parameter between the variant lymphocyte sample and the normal sample, the percent difference is as shown in Figure 3.7.

	NE		LY		MO		EO	
	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
V	-5.51%	-12.45%	5.69%	38.74%	-1.96%	27.56%	0.89%	13.30%
C	5.31%	2.52%	-0.92%	125.12%	3.88%	60.31%	8.38%	76.29%
S	10.99%	3.08%	-19.72%	35.16%	-4.42%	31.60%	1.18%	3.05%

Figure 3.7: Percentage Difference of Statistical Parameters (Variant Lymph Sample minus Normal Sample) Paying particular attention to the Lymph and Mono populations (since the Eo population does not contain a sufficient number of events to be considered statistically relevant), it is noticeable that the SD parameters for Volume, Conductivity and Scatter show the largest changes, with the means showing some changes as well. These changes points to the fact that the SD parameters derived from the Volume, Opacity, and Scatter of specific populations are useful and powerful features to detect abnormalities in population distributions, such as those shown in the previous example.

2. Case #2: Medically-Positive Variant Lymphocyte Sample

Another typical example of a sample having a medically-abnormal level of variant lymphocytes is shown below:

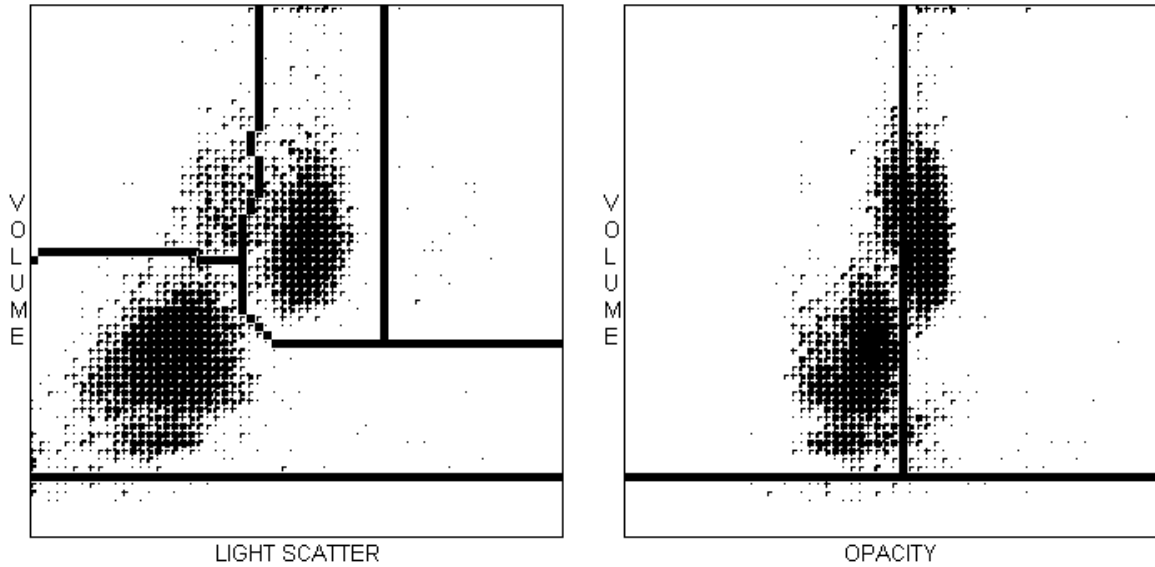


Figure 3.8: Blood Pattern with Abnormal Level of Variant Lymphocytes – Case #2

For this sample, The LH750 differential algorithm yielded the WBC percentage results given in Table 3.3.:

Table 3.3: LH750 Algorithm Reported WBC Percentages for Case #2

Lymph%	Mono%	Neutro%	Eo%	Baso%
62.49%	4.099%	32.62%	0.276%	0.512%

Likewise, the manual reference results are given in Table 3.4

Table 3.4: Manual Differential WBC Percentages for Case #2

Lymph	Mono	Neutro	Eo	Baso	Blast	Immature Gran	Immature Band	Variant Lymph	Plasma Cell
47.5%	0.5%	14.75%	1.75%	0%	0%	0%	13%	22.5%	0%

This sample is of primary importance not only because its manual variant lymphocytes percent is 22.5, well above the medically-accepted threshold to be considered abnormal, but because currently the LH750 differential algorithm did not detect this sample as having variant lymphocytes. Thus, it is currently considered a false negative. This is a type of pattern for which some improved detection is necessary in order to help resolve this type of misclassification occurrence.

3. Case #3: Medically-Negative Variant Lymphocyte Sample

The next sample shown is a primary example of a sample that should not mistakenly be flagged as having excessive variant lymphocytes. It is the pattern of a clinically normal sample, with a normal level of variant lymphocytes. The LH750 Variant Lymph detection algorithm does not detect this pattern as having excessive variant lymphocytes; thus it treats this sample in the correct fashion, as a True Negative. A well-designed classification method for variant lymphocytes should also be able to distinguish this pattern from a positive sample without difficulty. This pattern is shown below:

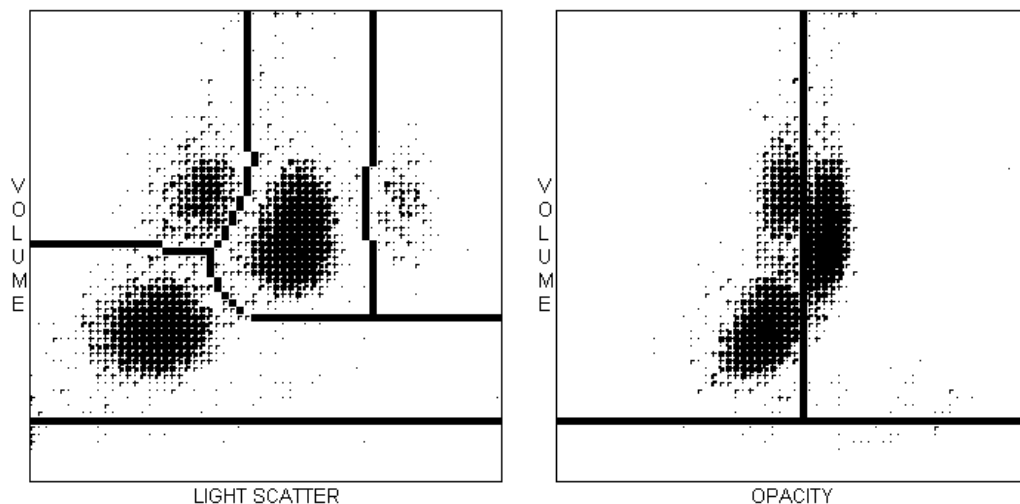


Figure 3.9: Blood Pattern with Normal Level of Variant Lymphocytes – Case #3

For this donor's blood, the LH750 Differential algorithm gives the following WBC percentages:

Table 3.5: LH750 Algorithm Reported WBC Percentages for Case #3

Lymph%	Mono%	Neutro%	Eo%	Baso%
38.102%	8.401%	50.754%	2.174%	0.568%

Also, the Manual Differential is given for this donor as follows:

Table 3.6: Manual Differential WBC Percentages for Case #3

Lymph	Mono	Neutro	Eo	Baso
37%	10%	49%	2%	0%

4. Case #4: Medically-Negative Variant Lymphocyte Sample

An example of an acute lymphoblastic leukemia is shown below. This sample has its hematology scatterplots shown in Figure 3.10.

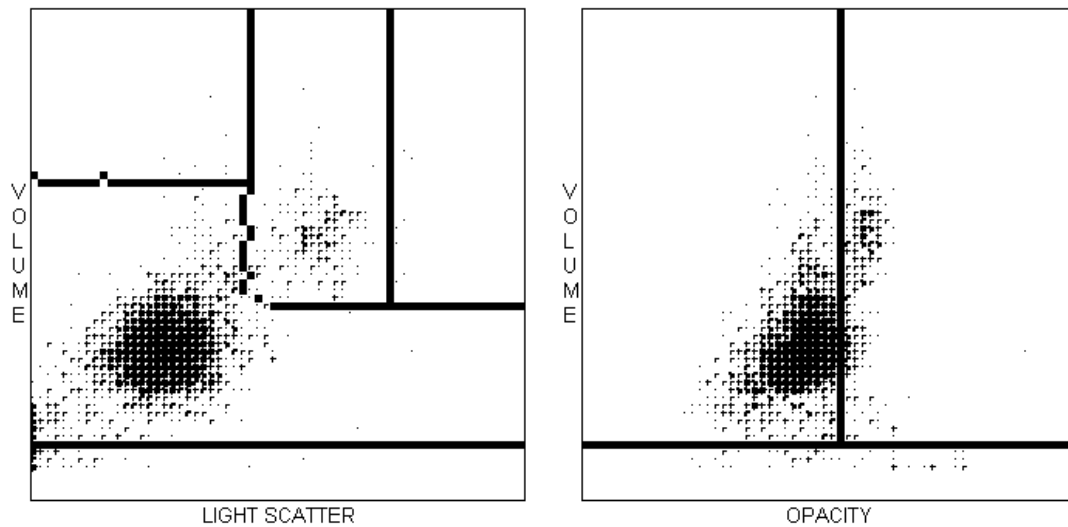


Figure 3.10: Blood Pattern with Blasts but Normal Level of Variant Lymphocytes – Case #4

The LH750 differential algorithm produced the following WBC percentages as shown in Table 3.7.

Table 3.7: LH750 Algorithm Reported WBC Percentages for Case #4

Lymph%	Mono%	Neutro%	Eo%	Baso%
91.989%	0.23%	7.485%	0.23%	0.066%

From the above figure, it is apparent that a predominance of lymphocytes exists in this sample, attesting that the disorder is of a lymphocytic origin. In this particular case, lymphocytes comprise approximately 92% of the WBC events on the scattergrams. However, it is important to note that from the point of view of a typical hematology analyzer with no specific monoclonal antibodies or markers, it is not easily apparent as to whether or not these are immature lymphocytes or mature lymphocytes as contained within the population shown above. The most obvious pattern feature is the predominance of lymphocytes (lymphocytosis), however, according to several medical texts and publications, the presence of lymphocytosis is not always a reliable feature, since it can sometimes be present, but not be malignant in nature. Especially in children, lymphocytoses often occur but is of benign origin. In adult patients, however, lymphocytosis is normally a warning sign of a serious disorder, such as leukemia or lymphoma [Caldwell and Lacombe 2000]. From the perspective of the WBC manual differential, the cellular percentage statistics as given in Table 3.8 were reported for this donor.

Table 3.8: Manual Differential WBC Percentages for Case #4

Lymph	Mono	Neutro	Eo	Baso	Blast	Immature Gran	Immature Band	Variant Lymph	Plasma Cell
81%	0%	8%	1%	0%	9%	0%	1%	0%	0%

According to information provided by the doctors for this patient, the Blasts present in this sample (roughly 9% as a fraction of total WBCs) were determined to be of lymphocytic origin and are B-cells (produced by patient's bone marrow). The importance of the B Blast cells in this sample are that they can cause a distortion of the shape of the lymphocyte population, since Blast cells can often have very different volume, light scatter, or opacity characteristics than normal lymphocytes.

However, automated hematological analyzers such as the Coulter LH750 series are not capable of separating immature cells from their mature counterparts. This is an unfortunate limitation of VCS technology. The best recourse that an automated hematology analyzer can take is to give a reliable indication to the user of the device that an abnormality has been detected in the pattern of the scattergram or in the measured statistics of the sample under consideration. Samples with a rather high proportion of immature lymphocytes (Blasts), such as the pattern from this donor, can sometimes satisfy the conditions of a variant lymphocyte classifier, thus causing a flag, inducing a false positive for variant lymphocytes. However, from a patient risk perspective, it is much safer to give a false positive detection than a false negative one, especially for a serious disorder such as acute lymphocytic leukemia.

3.4. Selection of Pattern Features for Variant Lymphocyte Sample Classification

As shown in the previous section, population statistics (percentages, mean, and standard deviation) are often used to delineate normal from abnormal patterns. For a large number of patterns, this approach can be effective.

In terms of classification of variant lymphocyte samples (medically abnormal versus medical normal levels), the following population statistical features are most often used:

- Lymphocyte Percentage
- Lymphocyte Volume Mean
- Lymphocyte Volume Standard Deviation
- Lymphocyte Opacity Mean
- Lymphocyte Opacity Standard Deviation
- Lymphocyte Light Scatter Mean
- Lymphocyte Light Scatter Standard Deviation

Commonly, to assess the classification “ability” of a population statistical parameter or feature to classify an abnormality of interest, a basic method employing histogram overlay plots is often used, where two 1-D histograms are shown. The first histogram plotted will represent the distribution of the feature using a set of data that is “positive” for the abnormality of interest, and the second histogram plotted will represent the distribution of the feature using a set of data that is “negative” for the abnormality of interest. An example is shown in Figure 3.11 demonstrating the feature “Lymphocyte Volume Standard Deviation” for samples of the two classes, “Medically Normal Levels of Variant Lymphocytes” and “Medically Abnormal Levels of Variant Lymphocytes.”

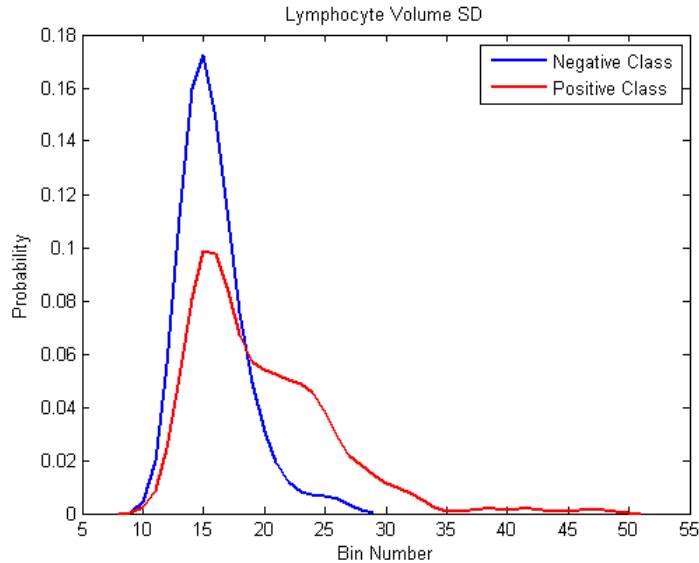


Figure 3.11: Histogram Overlay of Lymphocyte Volume SD

For this parameter, it is noted that the “Variant Lymphocyte Negative” samples, or the Negative Class, shown in blue, have a roughly-symmetric, almost Gaussian distribution with a mode at about bin 15. On the other hand, the “Variant Lymphocyte Positive” samples, or the Positive Class, shown in red, have a heavily-skewed, asymmetrical distribution, also with a mode at bin 15, but with a secondary peak near bin 25.

It should be noted that the degree of overlap between these two histograms is the important aspect to be considered here. The lower the degree of overlap between the classes in this feature space, the better the ability of this feature to offer class separability. For this particular feature, it is noticeable that the degree of overlap is very high, but there is a region (above bin 22) where these feature histograms have minimal overlap. This indicates that this feature has some value for separating these classes, but to use this feature alone for classification would incur a large classification error. One could, of course, choose different threshold levels, but one would not escape the problem that these sample types are not linearly separable in this 1D feature space (Lymph Volume SD).

For all six basic population statistical parameters (Lymph Volume Mean, Lymph Volume SD, Lymph Opacity Mean, Lymph Opacity SD, Lymph Light Scatter Mean, and Lymph Light Scatter SD), the different histogram overlays are shown in Figure 3.12.

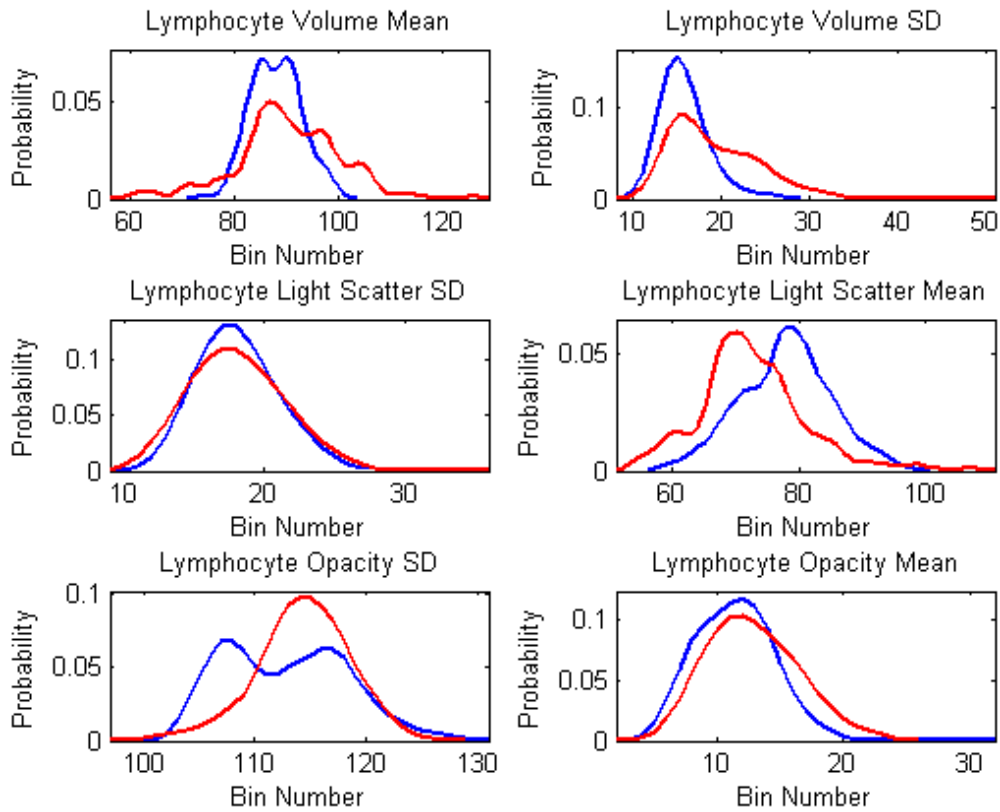


Figure 3.12: Histogram Overlays of Lymphocyte Population Statistical Parameters

It is apparent from examination of each of these histogram overlays that each lymphocyte population statistical parameter offers a varying degree of class separability. Some parameters offer little or no separation of the classes (Light Scatter SD and Opacity Mean) and some seem to offer a fair amount of separation. Therefore, the question arises as to what is a good method for numerically assessing the class separability that can be offered by a particular parameter, or feature, in seeking effective classification.

This again brings up the discussion of the set of topics that were introduced in Section 2.2.2 on ROC curve analysis. The ROC curve again can be useful, in the fact that it can be used to quantify the “degree of correct classification” at every potential threshold value of the parameter or variable used. It is conceptually equivalent to allowing the vertical line shown in Figure 3.13 (which represents a threshold value, or decision point) to vary from all bin values (from bin 5 to bin 55 in the example), and at each bin value evaluating the percentage of true positive patterns and the percentage of false negative patterns. The false positive fraction (FPF) and true positive fraction (TPF) values are then plotted as a set of values (X, Y) for every bin value used, and an ROC curve will be generated.

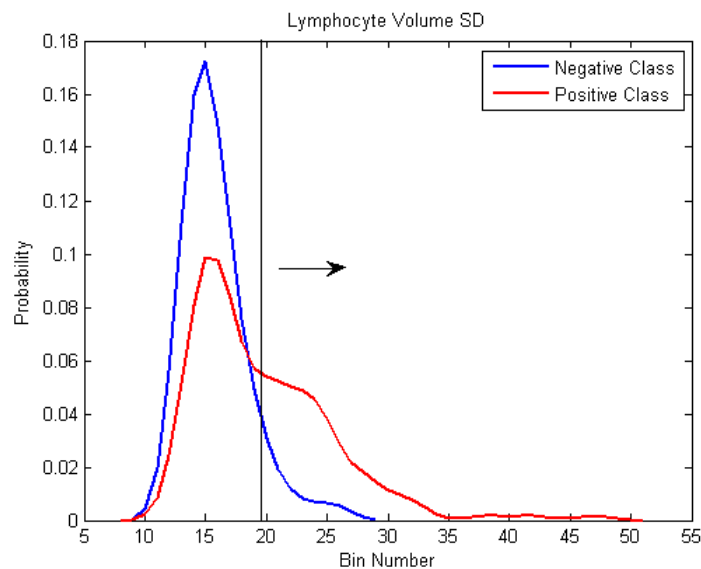


Figure 3.13: Histogram Overlay with "Sliding Decision Point Bar"

A well-known classification efficiency parameter that can be calculated for an ROC curve is the AUC, also known as the *area under the curve*; this parameter is quite useful because it serves to quantify the potential classification “value” or class separability that a feature of interest can offer. The AUC measurement is often normalized to the range

[0, 1], where an AUC value of 1 represents a perfect classification outcome. Typically, as a rule of thumb, classification variables with values of AUC at or below the value of 0.5 are considered random in their classification abilities and are normally discarded in lieu of better-chosen parameters [Tilbury et al 2000].

For the Lymphocyte Volume SD parameter shown, the following ROC curve is generated.

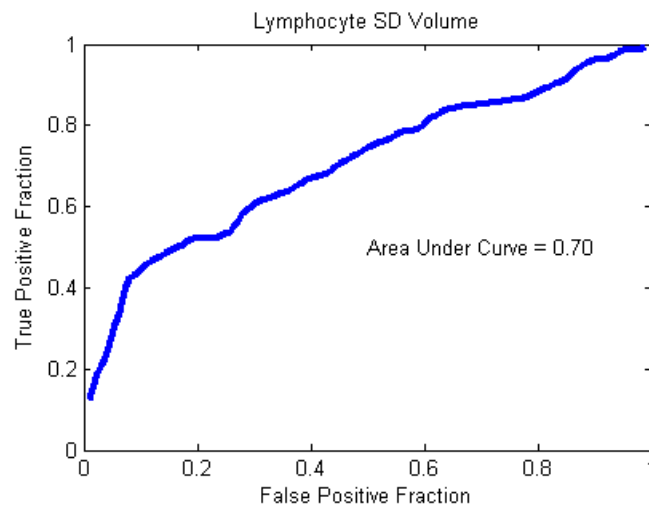


Figure 3.14: ROC Curve for Lymphocyte Volume SD

For this feature, the AUC value is 0.7, which indicates that it has some potential for classification of variant lymphocyte samples. Likewise, for all the 6 lymphocyte population statistical parameters, their respective ROC curves are shown in Figure 3.15.

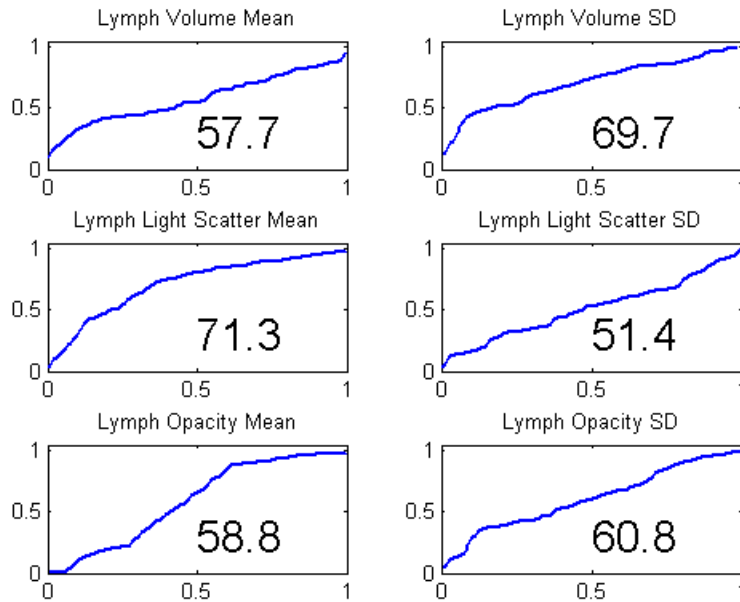


Figure 3.15: ROC Curves for all 6 Lymphocyte Statistical Parameters

3.5. Practical Difficulties With Variant Lymphocyte Classification Methods

As was discussed in Section 2.2.5, one of the major limitations of automated hematology methods is accurate and sensitive flagging of variant lymphocyte patterns, especially for donors that only express slightly elevated levels of variant lymphocytes. When the levels of variant lymphocytes begin to exceed 10% (as a proportion of total WBC cells), the hematological pattern presented by an automated instrument usually have the tendency to appear abnormal [Bessman 1986]. Pattern “abnormalities” due to the presence of variant lymphocytes can manifest themselves in many different ways, as was made evident through the previous case studies summarized in the previous section. However, a fairly common pattern manifestation that can result from excessive levels of variant lymphocytes is to see an obvious increase in the width or elongation of the lymphocyte population. This effect shown in Figure 3.16 is the relationship between the Manual

Variant Lymphocyte Percent (obtained by light-microscopic methods) and the Lymphocyte Volume SD.

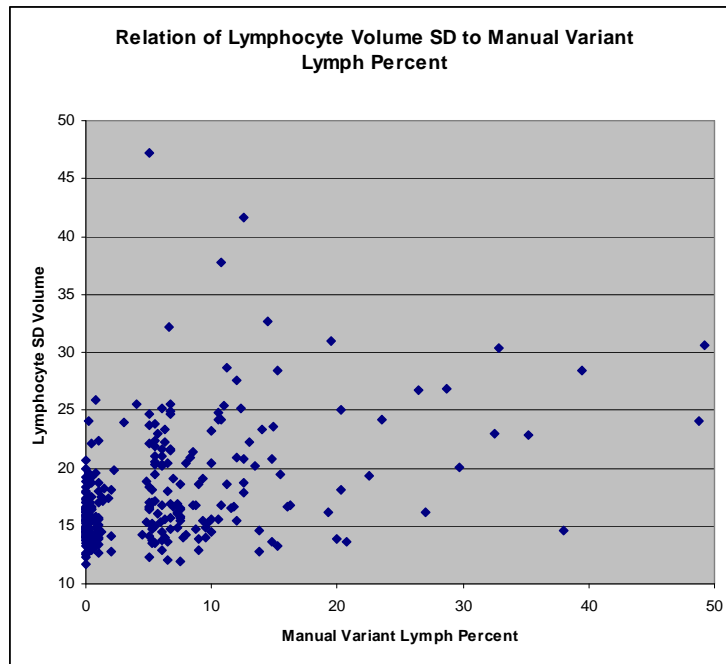


Figure 3.16: Relation of Lymphocyte Volume SD to Manual Variant Lymphocyte Percent

As seen in the above figure, there is a tendency for the standard deviation of the volume of the lymphocyte population to grow as the proportion of variant lymphocytes increases. But, from previous analyses this feature is not sufficient to build an efficient classifier, since some samples have very large variant lymph percents yet do not express a large standard deviation in their volumes. This assertion has an obvious medical relevance, since some types of lymphocytes that are termed “variant” and hence abnormal by clinicians and biologists are merely lymphocytes that appear larger in physical size (volume) than so-called normal lymphocytes when viewed on microscopic slides [Turgeon 2005] [van der Meer et al 2007].

Also, many samples exhibiting medically-abnormal levels of variant lymphocytes tend to express an obvious lymphocytosis, which is an increase in the overall proportion of lymphocytes. This is due to the fact that many diseases and disorders that can cause variant lymphocytes to appear in the blood tend to be lymphoproliferative, that is, they have a predominant effect on WBC cells from the lymphocytic lineage [Caldwell and Lacombe 2000] [Bessman 1986] [Turgeon 2005] .

For this reason, one of the most useful features in the classification of lymphocyte abnormalities is the percentage of lymphocytes relative to the whole WBC population. This parameter of course has an obvious medical relevance since many lymphocytic disorders (especially chronic lymphocytic leukemia) are known to express an obvious lymphocytosis (abnormally high proportion of lymphocytes as a fraction of total WBC cells) as a primary feature [Caldwell and Lacombe 2000]. The lymphocyte percent ROC curve is illustrated in Figure 3.17.

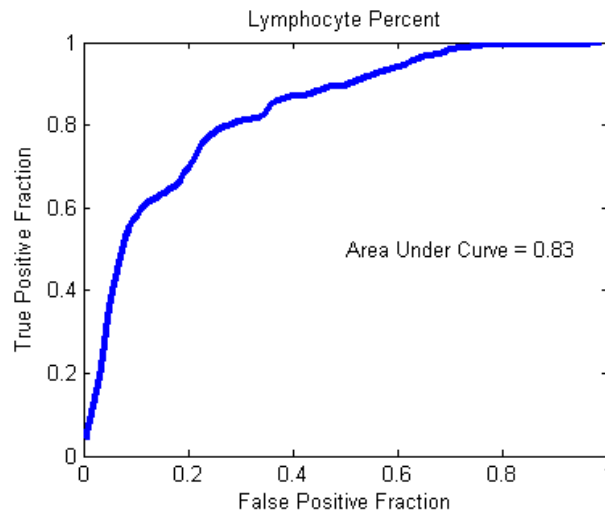


Figure 3.17: ROC Curve of Lymphocyte Percent Parameter

So far, by using the population percentage and basic second-order statistics of the lymphocyte population itself, the AUC values obtained are as shown in Table 3.9. In the table, the following abbreviations are used (V = Volume, LS = Light Scatter, OP = Opacity). Note that these values are sorted in decreasing order.

Table 3.9: ROC Performance Metrics of Common Lymphocyte Population Statistical Parameters

Parameter	Area Under Curve	Optimal Sensitivity	Optimal Specificity
Lymph Percent	83.0%	78.6%	75.0%
Lymph LS Mean	71.3%	66.0%	67.8%
Lymph V SD	70.1%	62.1%	66.9%
Lymph OP SD	60.5%	55.0%	56.9%
Lymph OP Mean	58.5%	52.9%	57.5%
Lymph V Mean	57.5%	54.3%	53.7%
Lymph LS SD	51.0%	47.1%	48.1%

From these ROC analyses and other medically-based rationales, the features shown above seem to be logical choices with which to design a classifier of variant lymphocyte samples.

3.6. Theory of Fisher Linear Discriminants

With the multitude of potential features available for pattern discrimination, it is necessary to define a quantitative method with which to assess the effectiveness of a given parameter. One popular method is the Fisher linear discriminant. Discriminant analysis seeks directions in the feature space that are efficient for discerning differences between pairs or groups of patterns. This idea is an elegant one, as it considers the problem of projecting data from its multi-dimensional space onto a line, thus providing a method of dimensionality reduction. Of course, it is possible that even if the sample data samples form well-separated, compact clusters in multi-dimensional space, simply

projecting the data onto an arbitrary line would probably produce a confused mixture of samples and thus produce a poor recognition performance. For this reason, it is necessary to move the line around mathematically to eventually find an orientation direction for which the projected sample data are well separated. This is exactly the goal of discriminant analysis [Duda, Hart, and Stork 2001].

Suppose that we have a set of n D-dimensional data samples x_1, x_2, \dots, x_n , where n_1 is the subset of samples D_1 that is labeled w_1 , and n_2 is a subset of samples D_2 that is labeled w_2 . If a linear combination of the components of x is formed, the following scalar dot product is obtained:

$$y = w^T x \quad (3.1)$$

And a set of n samples y_1, y_2, \dots, y_n which have been divided into subsets Y_1 and Y_2 . If $\|w\| = 1$, each y_i is the projection of the corresponding x_i onto a line in the direction of w . Thus, the direction of w is important and determining the appropriate w that provides the best separation between the subsets w_1 and w_2 is the ultimate goal.

A measure of the separation between the projected points is the difference of the sample means. If m_i is the D-dimensional sample mean given by:

$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x \quad (3.2)$$

Then the sample mean for the projected points is given by:

$$\begin{aligned} \tilde{m}_i &= \frac{1}{n_i} \sum_{y \in Y_i} y \\ &= \frac{1}{n_i} \sum_{x \in D_i} w^T x = w^T m_i \end{aligned} \quad (3.3)$$

And is simply the projection of m_i . And thus, it follows that the distance between the projected means is:

$$\left| \tilde{m}_1 - \tilde{m}_2 \right| = \left| w^T (m_1 - m_2) \right| \quad (3.4)$$

This difference can be scaled by simply changing the value of w . However, if the goal is ultimately to obtain good separation of the projected data, the difference between the means “should be large relative to the standard deviations for each class” [Duda, Hart, and Stork 2001]. A popular measure that is proportional to the standard deviation of projected samples is known as the scatter and is defined as:

$$s_i^2 = \sum_{y \in Y_i} (y - m_i)^2 \quad (3.5)$$

The quantity $\frac{1}{n}(s_1^2 + s_2^2)$ is an estimate of the variance of the pooled data, and the quantity $s_1^2 + s_2^2$ is called the within-class scatter of the projected samples. Using the combination of the differences of the means and the within-class scatter measurement, the Fisher Linear Discriminant defines the following criterion function as a function of w :

$$J(w) = \frac{\left| \tilde{m}_1 - \tilde{m}_2 \right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \quad (3.6)$$

This last equation is significant because it shows that the w value that maximizes the criterion function $J(w)$ leads to the best separation between the two projected sets.

3.7. Use of Fisher Linear Discriminants for Feature Selection and Evaluation

In the event that none of these features can separate the classes of data on their own, it becomes necessary to find multi-dimensional descriptors that potentially can do the task

more effectively. One popular method of linearly combining a number of 1-dimensional features to create a multi-dimensional descriptor (or classifier) is through the method of Fisher linear discriminants as will be detailed in Section 3.8.5.

3.8. Multi-Dimensional Feature Extraction Method for Improved Pattern Recognition

A key contribution of this dissertation is to propose the extraction of a set of features that are more descriptive of hematological data patterns than simple 1-dimensional population statistical features that may not always be representative of the abnormality of the pattern. This is why the notion of “shape descriptors” has become such a commonly used term in more recent pattern recognition approaches, especially in research areas whose primary focus is on complex image-processing methods, such as medical imaging, face recognition and content-based image retrieval [Kim and Kim 2000].

The inherent nature of hematological data is that it is multi-dimensional, since several measurements are simultaneously collected for each analyzed blood cell. The problems arise when one attempts to graphically illustrate clusters or populations of data in their native multi-dimensional states. Even though many methods have been proposed over the years, the most popular surface rendering method is still done using the well-known “Marching Cubes” algorithm [Lorenson and Cline 1987]. It basically takes a multi-dimensional dataset, and extracts from it a triangular mesh that is representative of the outer surface of the data cluster. Using the Marching Cubes algorithm along with the MATLAB isosurface command, an example of Normal donor’s hematological data is illustrated in Figure 3.18 using surface rendering.

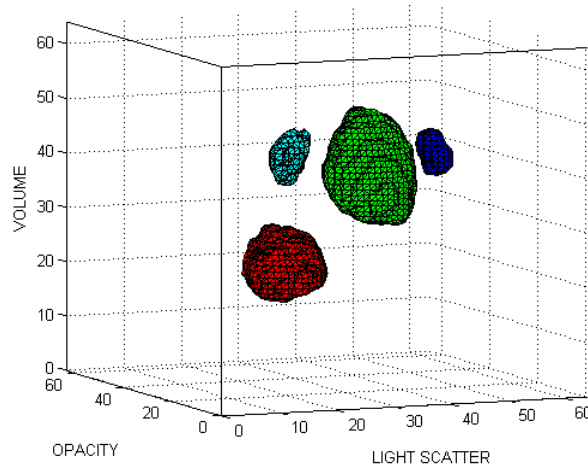


Figure 3.18: Surface Rendering of LH750 WBC Differential Data of a Normal Donor

From the 3D surface rendering of an LH750 hematological data pattern, it should be noted that each population (Lymphocytes shown in Red, Monocytes shown in Cyan, Neutrophils shown in Green, and Eosinophils shown in Blue), has its own unique mean location, variance and orientation (pose). It may also be noted that the pose or directional orientation of each population is not always necessarily orthogonal to the measurement directions of Volume, Opacity, and Light Scatter. For this reason, describing the pose of the 3D populations with vectors which are free to conform to the structure of the data might be advantageous. In the realm of multivariate statistics, this is a commonly encountered problem, and a classical method is often employed for this type of analysis which is known as the Karhunen-Loeve transform, or Principal Component transform.

3.8.1. Mathematical Aspects of the Principal Component Transform

Given that a dataset consists of n d -dimensional samples x_1, x_2, \dots, x_n , the question often arises as to how to represent all of the samples concisely as a single vector x_0 . To be more specific, it may be desired to determine a vector x_0 such that the “sum of squared distances between x_0 and all the various samples x_k is as small as possible” [Duda et al

2000]. Formulating this problem mathematically can be accomplished by defining the following squared-error criterion function:

$$J_0(x_0) = \sum_{k=1}^n \|x_0 - x_k\|^2 \quad (3.7)$$

After defining this functional, the value of x_0 that maximizes J_0 is obviously sought. It can be easily shown that the solution to this problem is given by $x_0 = \mathbf{m}$, where \mathbf{m} is the sample mean. However, this result is not interesting because the sample mean is a “zero-dimensional representation of the data set” that does not contain any information about the variation in the data. On the other hand, a much more useful one-dimensional representation of the data is the one that is created by projecting the data onto a line that runs through the sample mean [Jolliffe 1986] [Duda et al 2000]. Let e be a unit vector in the direction of the line, with its equation written as:

$$x = m + ae \quad (3.8)$$

The scalar value a (which can be any real number) corresponds to the distance of any point x from the mean m . If we represent x_k by $m+a_k e$, a set of optimal coefficients a_k can be found by minimizing the squared error criterion function:

$$\begin{aligned} J_1(a_1, \dots, a_n, e) &= \sum_{k=1}^n \|(m + a_k e) - x_k\|^2 = \sum_{k=1}^n \|a_k e - (x_k - m)\|^2 \\ &= \sum_{k=1}^n a_k^2 \|e\|^2 - 2 \sum_{k=1}^n a_k e^t (x_k - m) + \sum_{k=1}^n \|x_k - m\|^2 \end{aligned} \quad (3.9)$$

Noting that the magnitude of e is 1 (since it is a unit vector), and partially differentiating equation 3.9 with respect to a_k , and setting the derivative of the obtained function to zero, the following equation is obtained:

$$a_k = e^t (x_k - m) \quad (3.10)$$

This result is important because it states that a least-squares solution can be obtained by projecting the vector x_k onto the line e that passes through the sample mean. However, this formulation does not help us find the best direction of e for the line. To find this solution, the notion of the sample scatter matrix needs to be introduced and then incorporated into the problem. Thus, the scatter matrix, S , is defined as:

$$S = \sum_{k=1}^n (x_k - m)(x_k - m)^t \quad (3.11)$$

If equation 3.6 is substituted into equation 3.5, the following result is obtained:

$$\begin{aligned} J_1(e) &= \sum_{k=1}^n a_k^2 - 2 \sum_{k=1}^n a_k^2 + \sum_{k=1}^n \|x_k - m\|^2 \\ &= -\sum_{k=1}^n [e^t (x_k - m)]^2 + \sum_{k=1}^n \|x_k - m\|^2 \\ &= -e^t S e + \sum_{k=1}^n \|x_k - m\|^2 \end{aligned} \quad (3.12)$$

It may be noted that the vector e that minimizes J_1 also maximizes $e^t S e$. Using the method of Lagrange multipliers to maximize $e^t S e$ subject to the constraint that $\|e\| = 1$. If we let λ be the undetermined multiplier, then the following equation must be differentiated with respect to e :

$$u = e^t S e - \lambda(e^t e - 1) \quad (3.13)$$

After differentiating this equation, the following equation is obtained:

$$\frac{\partial u}{\partial e} = 2S e - 2\lambda e \quad (3.14)$$

And upon setting this differential equation equal to zero, an important result emerges since it is immediately noticed that e must be an eigenvector of the scatter matrix:

$$S e = \lambda e \quad (3.15)$$

Basically, because $e^t S e = \lambda e^t e = \lambda$, it follows that to maximize $e^t S e$, the eigenvector corresponding to the largest eigenvalue of the scatter matrix is to be selected. Thus, in order to find the “best one-dimensional projection of the data in a least sum-of-squared-error sense”, projecting the data onto a line through the sample mean in the direction of the eigenvector of the scatter matrix having the largest eigenvalue performs this task [Jolliffe 1986]. This method, of course, can be readily extended to multiple dimensions. Extending from a one-dimensional projection to a d-dimensional projection changes equation 3.8 to the following:

$$x = m + \sum_{i=1}^d a_i e_i \quad (3.16)$$

with the criterion function redefined as follows:

$$J_d = \sum_{k=1}^n \left\| \left(m + \sum_{i=1}^d a_{ki} e_i \right) - x_k \right\|^2 \quad (3.17)$$

This function is thus minimized when the vectors e_1, \dots, e_d are the d eigenvectors of the scatter matrix having the largest eigenvalues. Since the scatter matrix is real and symmetric, these eigenvectors are orthogonal. The coefficients a_i in equation 3.16 are the principal components of the feature vector x in that basis [Jolliffe 1986] [Jackson 1991] [Duda, Hart, and Stork 2001].

3.8.2. Principal Component Transform for 3D Variance and Pose Representation

To apply method of Principal Component Analysis to hematological data from the LH750 analyzer in a meaningful way, the raw data points (Volume, Light Scatter, and Opacity measurements) corresponding to each individual population should first be obtained, and then arranged into arrays similar to the one represented in Table 3.10.

Table 3.10: Raw Event Array for a single hematological population collected with an LH750 analyzer

Population Event #	Volume Measurement (X)	Opacity Measurement (Y)	Light Scatter Measurement (Z)
1	45	54	34
2	52	43	37
.....
N	38	58	36

Once the arrays of data for each hematological population have been obtained, extraction of the eigensystem of each data array proceeds with the standard PCA formulations [Gonzalez and Woods 1992].

Since the covariance matrix C_x is real and symmetric, finding the set of \mathbf{n} orthonormal eigenvectors is always possible. The eigensystem of the covariance matrix is then found using traditional methods. For the 3-dimensional case, the characteristic equation would be formulated using a 3x3 covariance matrix in search of the roots λ_i , which in turn will determine the eigenvectors corresponding to the the three directions of largest variance.

As an illustrative example of the usage of the eigensystem in 3D has been extracted for a single population and is displayed in Figure 3.19.

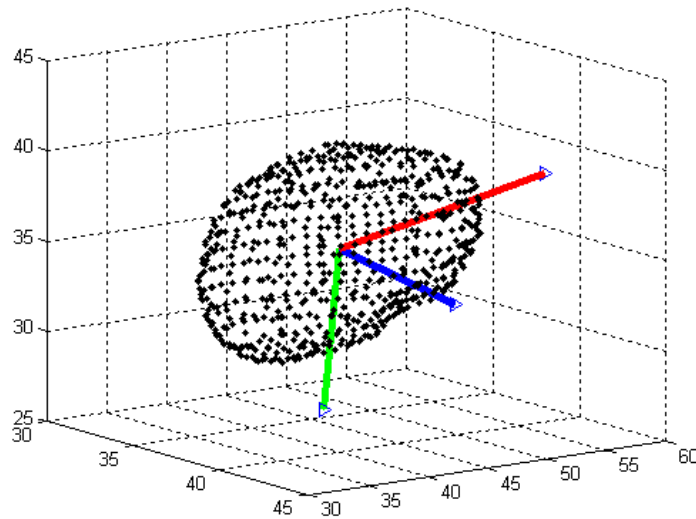


Figure 3.19: Eigensystem pose and variance descriptors for a single WBC population

3.8.3. Extracting Eigensystem-Based Population Descriptors

Since it is fairly well known that the standard principal component analysis methods based on the extraction of the eigensystem of the covariance methods are based on minimization of L2-norms, or squared-error criteria, the method contains inherent sensitivity to outlier points. This happens because the “effect of outliers with large norm values is exaggerated by the use of the L2-norm” [Kwak 2008] [Ding et al 2006]. For these reasons, approaches such as Weighted-PCA, L1-Norm PCA and other methods of incorporating robustness through reduction of the effects of outlier data on the calculation of the eigenvectors and/or the covariance matrix have become popular in recent years [Kriegel, Kröger, Schubert, and Zimek 2008]. It should be noted, however, that the use of some of these methods, such as those based on calculating “robust” estimates of mean, median, and variance from large datasets, can be unwieldy in terms of computational efficiency. Since it is desired that the eigensystem extraction be done with minimal computational requirements, care was taken to implement a method for extracting the

PCA that maintains some of the “outlier” reduction strategies of robust estimation methods while still providing the final result in an acceptable processing time.

3.8.4. Proposed Extraction Method for Eigensystem-Based Cluster Descriptors

The flowchart given in Figure 3.20 shows the processing steps used to extract the eigensystem for the hematological data patterns considered in this dissertation. It shows how the set of 52 eigensystem features (which are representative of the 4 major WBC populations) are produced from the raw data that is given in the form of an $N \times 3$ data array:

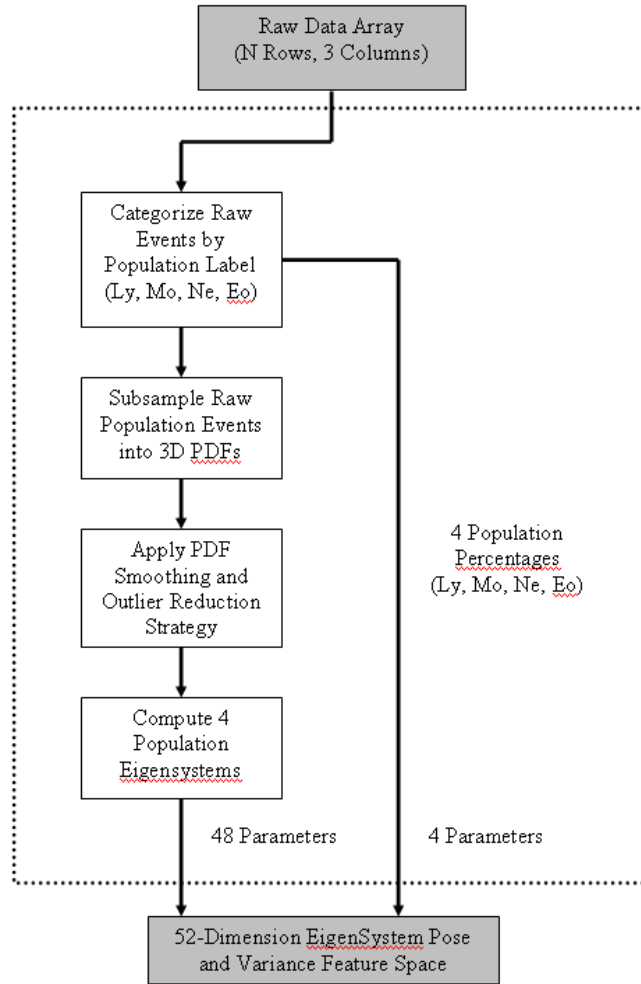


Figure 3.20: Flow Chart Depicting Process of Extraction of Eigensystem-Based Descriptor Matrix

According to the flowchart shown above, the process of extraction of the 52-parameter descriptor matrix makes some important considerations:

- **Data Resolution Subsampling:** The resolution of each measurement (Volume, Light Scatter, Opacity) is initially 256 channels, or 8 bits. Since it is desired to accumulate the data into 3D histograms, production of 256x256x256 channel histograms would be required; dealing with histograms of this magnitude could become computationally unwieldy. Thus, it was experimentally determined that 3D histograms with 32 channels (1/8 the number of channels) could give an accurate representation of the hematological pattern information with the added benefit of taking a factor of $(8 \times 8 \times 8 = 8^3)$ less computation time to produce and process mathematically.
- **Data Smoothing and Outlier Reduction:** Each population contains residual noise and events that are far from the cluster body. These events are normally considered statistical outliers and their effects first reduced by smoothing the data with Gaussian filtering in the spatial domain. In addition to smoothing, which does not remove completely the outlier points; morphological operations such as erosion and dilation are used, in conjunction with intensity thresholding in order to reduce the effects of noise and outlier points to an acceptable level.

3.8.5. Comparison of Features Using Fisher Linear Discriminant Analysis

After potential population features have been evaluated according to their ROC curve individually, and it was determined that no single feature can give the desired level of accuracy for the two-class classification problem at hand, a method of incorporating them into a multidimensional class separation framework is considered the next logical step.

Based on previous discussions, a number of feature values x could be combined in a weighted linear fashion (through the use of a set of weights w) into a single function y using the following equation:

$$y = w^t x \quad (3.18)$$

Assuming a two-class classification problem, the optimal value of w can be found using the following equation:

$$w = S_w^{-1}(m_1 - m_2) \quad (3.19)$$

In the following equation, m_1 and m_2 are the sample means of data points in class 1 and class 2, respectively. Likewise, S_w is known as the “within-class scatter matrix” and is found as follows:

$$S_w = S_1 + S_2 \quad (3.20)$$

S_1 and S_2 are the scatter matrices of the data points in class 1 and class 2, respectively. The scatter matrices S_i are calculated for the data points in each class as follows:

$$S_i = \sum_{x \in D_i} (x - m_i)(x - m_i)^t \quad (3.21)$$

In this case, finding the optimal value of w in the least-sum-of-squares sense is equivalent to finding the direction of best separation between the two classes of data. This property of the Fisher linear discriminant is then exploited to find the direction of best separation for any chosen number of potential features.

Using this method, the 7 lymphocyte population features shown in Table 3.9 were combined into a single feature. Using the linear discriminant method, the classification performance when using this feature gives the ROC curve shown in Figure 3.21.

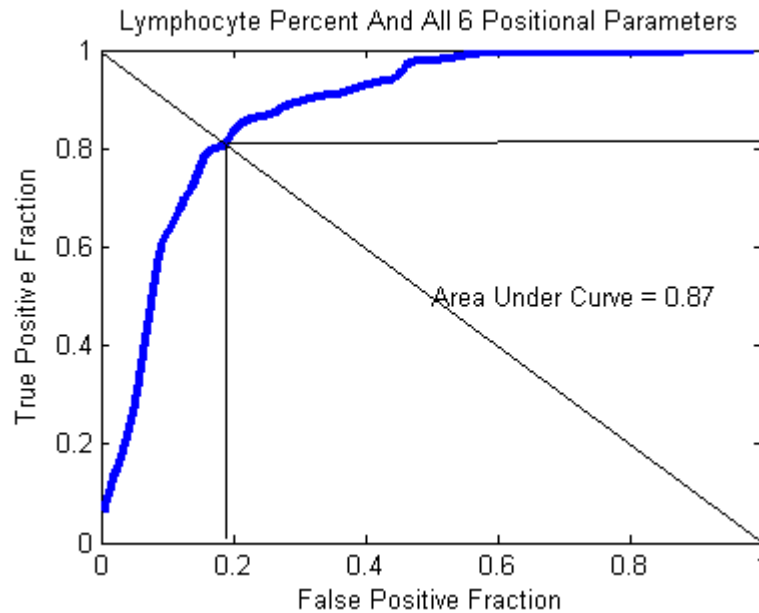


Figure 3.21: ROC Curve of Linear Classifier Using 7 Lymphocyte Statistical Parameters

With an AUC value of 0.87, and using the optimal closest point (the intersection point of the ROC curve with a diagonal line drawn from point (0,1) to (1,0)), gives a specificity of 81.5% and sensitivity of 80.5%. It can be seen that the accuracy of a classifier built with only these parameters is not likely to approach 100%, which is of course desired but not often attainable. For these reasons, it thus becomes necessary to resort to selection of feature parameters that have better ability to discriminate between the two classes of data. From the case studies that contained the variant lymphocyte cell type, most often from donors diagnosed with variations of lymphocytic leukemia, infectious mononucleosis, or lymphoma, it should be noted that these disorders had, of course, an obvious effect on the lymphocyte population, but also that other WBC populations were often affected. For instance, for cases with lymphocytosis (a predominance of lymphocytes as a fraction of total WBC cells), obviously, the proportions of other WBC cell types (Monocytes, Neutrophils, and Eosinophils) may have been different than their normally-expected

levels. Likewise, the positional statistics of these other WBC populations would likely be different than those from normal blood donors. Thus, perhaps, it is necessary to use statistics from all of the WBC populations to obtain improved classification accuracy for the variant lymphocyte detection method. This assertion has medical relevance, since many types of WBC cells are often produced and/or activated by the human immune system upon its recognition of a potential pathogen [Turgeon 2005].

It can be shown that the eigensystem-based population descriptors do, in fact, augment the separability of the two data classes, Variant Lymphocyte Negative, and Variant Lymphocyte Positive. The Fisher linear discriminant is used to combine 7 features. For the first case, the 7 standard features of the lymphocyte population that were shown in Table 3.9 are used to build a single feature. For the second case, three of the best features from the eigensystem of the lymphocyte population are used in place of the features Lymph Volume SD, Lymph Light Scatter SD, and Lymph Opacity SD and again, a single feature composed of 7 variables is made. These results are shown in Figure 3.22.

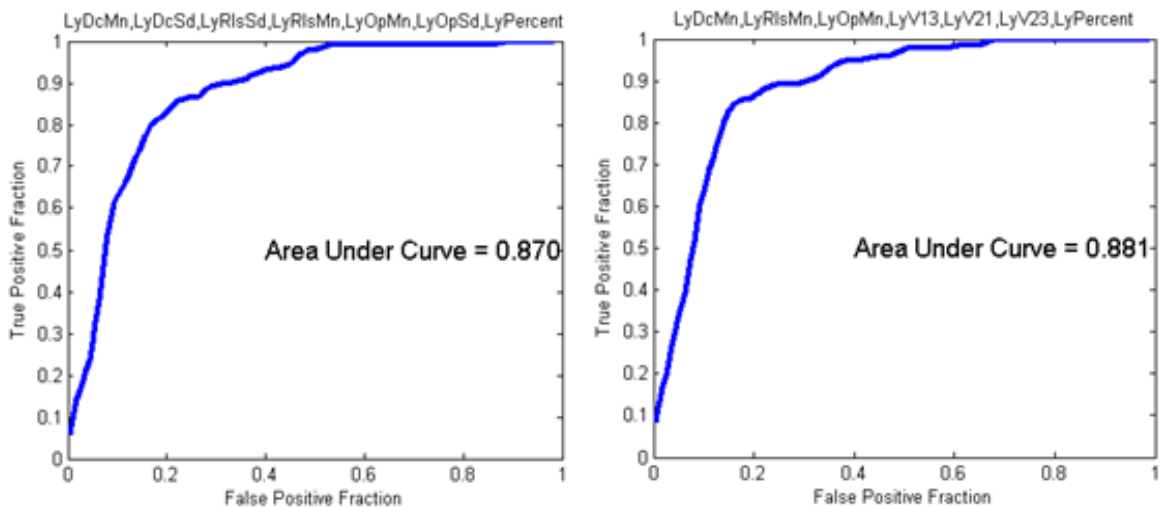


Figure 3.22: Comparison of ROC Curves of Two Linear Classifiers

3.8.6. Retrospective

From the results obtained, it is observed that the addition of the best lymphocyte eigensystem values (selected according to their AUC values) does improve the classification accuracy – this is evidenced by the slight increase of AUC value when comparing the classifier on the right to the one on the left.

From previous discussions on ROC curve theory, it is known that the eigensystem with the highest AUC scores can theoretically offer the best class separability. The top 10 parameters (sorted according to their AUC values) are shown in Table 3.11.

Table 3.11: Area-Under-The-Curve Values for the 10 Best Eigensystem Parameters

Parameter Name	AUC Value	Parameter Name	AUC Value
Ly Percent	0.842	Ne E1	0.737
Ne Percent	0.807	Ly V SD	0.712
Ly E1	0.774	Ly LS Mean	0.710
Ly E2	0.764	Ly V13	0.708
Ly E3	0.745	Ly V11	0.699

These outcomes also confirm some of the assumptions made in previous sections. Firstly, that some of the lymphocyte population eigensystem components have relatively high classification values, namely, the largest eigenvalue of the lymphocyte population for instance, also known as LyE1, has an AUC equal to 0.774, which is superior to the parameter Lymphocyte Volume SD, which was the previous best standard deviation parameter of the lymphocyte population – which had an AUC of 0.712. Also, it was fairly surprising that the Neutrophil population statistics (both the population percentage itself and some of its Eigen components) were seen to have very formidable classification

abilities. In fact, the Neutrophil percent was noted to have comparable classification abilities to that of the Lymphocyte percent, for this classification problem.

If we again try to use subsets of the parameter space (selecting N-tuples of the best individual parameters by goodness of their AUC values), the following classification accuracies as shown in Figure 3.23 are determined for N = 3, 5, 7, 10, 15, 20, 25, 30, 35, 40, 45, 50, and all 52 parameters, respectively.

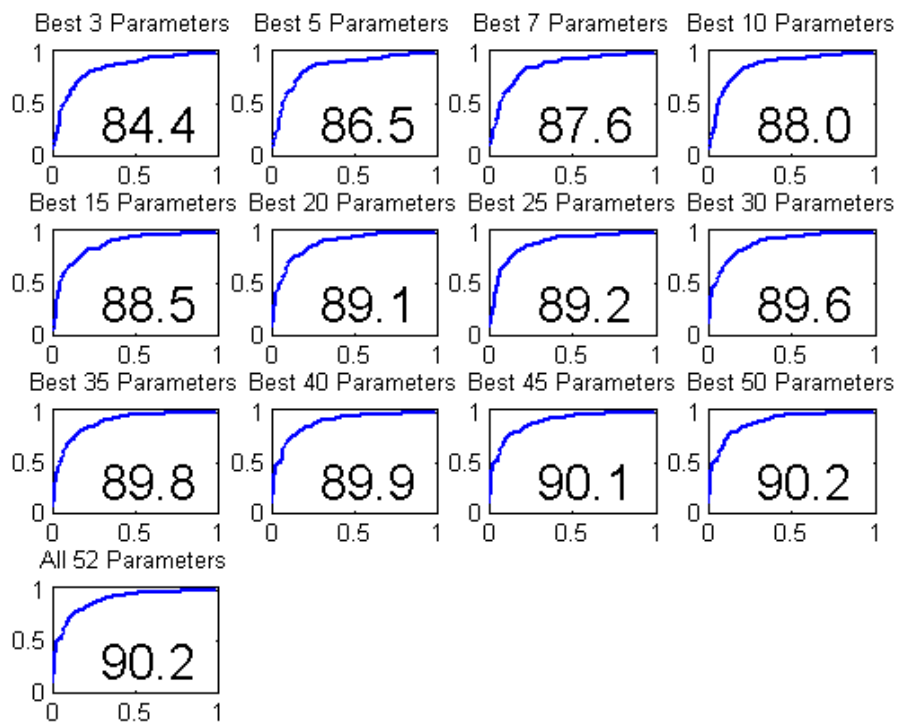


Figure 3.23: AUC Values for Classifiers Constructed from N-Tuples of Best Eigensystem Parameters

In the above figure, it can be observed that the best seven parameters of the Eigen component plus population percents classifier give slightly better results to the 7 parameters of the lymphocyte population alone (AUC = 0.876 versus 0.87). This result is

not surprising, since other features with better ability to separate the data classes are being added to the classifier, thus making a more realistic model of the disorder at hand. Upon adding more parameters to the classifier, naturally the classification accuracy is expected to increase. This is evident by seeing the progressively higher accuracy that is attainable by using a higher-dimensional feature space by increasing the number of parameters that have classification value. It can also be noted that the degree of increase of accuracy begins to become negligible as the number of parameters increases indefinitely. This is an example of the “law of diminishing returns”, as applied to pattern recognition. Thus, at some point, it may be sensible to perform a feature reduction (from the maximum of 52) through a sensitivity analysis to pare down the number of input features without incurring significant degradation to the overall accuracy of the classifier. However, it is clear that the performance will not degrade by adding more features, so reducing their number is not of great concern.

4. APPLICATION OF MACHINE LEARNING TO THE DETECTION OF VARIANT LYMPHOCYTE DATA PATTERNS

4.1. Objectives

The intent of this application is to seek two primary aims: (1) a high sensitivity (i.e., minimum number of false negatives), and (2) a high specificity (i.e., minimum number of false positives). A stopping condition for delineating Variant Lymph Positive from Variant Lymph Negative files assumes therefore the highest accuracy attainable will ensure that both primary aims are concurrently met.

Variant lymphocytes have historically been very difficult to detect reliably using conventional features extracted from hematological data patterns. Since this type of pattern detection that is done by the software algorithms of automated hematology analyzers must deliver its results quickly (typically in less than 5 seconds), a usable implementation can only be realized through a reliable and computationally efficient feature extraction and classification paradigm.

Over the past decade, several automated hematological detection paradigms focusing mainly on disease classification using extracted statistical population features have been reported with different degrees of success and inherent challenges. These studies included the use of rule-based linear classifiers, and application of neural networks [Zong et al 2005].

To be considered robust, a hematological pattern detection algorithm should then be sufficiently sensitive and specific, and any detection paradigm designed within it will need to capture the main features characterizing the abnormal blood pattern that differentiate it from a normal blood pattern. It has been demonstrated that there are often

inherent shape differences in individual WBC populations between samples that are abnormal and those that are normal. Namely, for clinically abnormal samples, the multi-dimensional clusters of data points representing each WBC population will generally show discernible differences in mean location, size, spread, or orientation (pose) when compared with identical populations from clinically normal donors. These empirical facts served as the foundation of this research endeavor.

In this chapter, the role of robust statistical operators that describe shape, spread, and pose of each population cluster are explored in order to develop a reliable real-time abnormal pattern detection algorithm for hematological data collected by the LH750 automated analyzer. The proposed method is based on extracting the set of orthogonal, shift-and-rotation-invariant statistical features derived from the eigensystem of each population cluster in three-dimensional space, as was described in detail in Chapter 3, and then applying this set of descriptive features to classify patterns that are indicative of the clinically-abnormal levels of variant lymphocytes. Since this feature space is multi-dimensional and complex, maximal utilization of its underlying descriptive power is most easily achieved through the use of machine learning paradigms. Thus, toward this end, an appropriate ANN architecture is established, with the parameters of this feature space as its inputs, and a training procedure is implemented to confront the complex nature of the classification problem at hand. The performance of the algorithm, which was evaluated by means of the ROC terminology, relied on two objectives: (1) establishing a decision space most suitable for variant lymphocyte pattern data classification, and (2) implementing an ANN that is trained to generate the weights for the highest classification accuracy possible. The proposed method looks at all Variant Lymphocyte Positive and

Variant Lymphocyte Negative files together with the purpose of creating an inter-patient classifier that would be applied irrespective of the particular patient under test.

4.2. Experimental Setup

The experiments were conducted using the same dataset of 300 donors that was introduced in Chapter 3. The data used in each one of the methods described in this study was obtained from a significant sample of patients who demonstrated varying levels of variant lymphocytes in their blood. Each of the hematological blood data was acquired using an LH750 Automated Hematology Analyzer operating in the differential mode of operation. Likewise, a trained hematologist performed a 200-cell Manual WBC differential count for each patient used in this study.

4.3. Dimensions of Feature Space

Assuming that the data pre-processing and feature extraction steps discussed in Section 3.8 have already been applied to each data file, a 13-dimensional representation for each of the 4 main WBC populations (where each population is represented by its 3 eigenvectors, 3 eigenvalues, and its population percent) is produced. Each eigenvector consists of 3 vector components (i, j, k), thus making 9 parameters. Thus, the proposed list of 52 potential features to be used in the classification of variant lymphocyte samples is as given in Table 4.1.

Table 4.1: Eigensystem Orientation and Variance Parameters for 4 WBC Populations

Population	Eigen Vectors	Eigen Values	Population Percents	Total Parameters
Lymphocyte	VL1, VL2, VL3	EL1, EL2, EL3	Ly%	13
Monocyte	VM1, VM2, VM3	EM1, EM2, EM3	Mo%	13
Neutrophil	VN1, VN2, VN3	EN1, EN2, EN3	Ne%	13
Eosinophil	VE1, VE2, VE3	EE1, EE2, EE3	Eo%	13

4.4. Appropriate Artificial Neural Network Design

Since multilayer feedforward neural networks, known as Perceptron are relatively simple and have been used successfully in a multitude of classification tasks, even for data that are not easily separable using conventional means (such as through linear classifiers), this type of ANN network will be applied to this classification problem using the 52-dimensional parameter space introduced in the previous section. Another advantage of this type of network is that it is easy to train by use of the much-cited backpropagation algorithm, which is based on successive optimization by backpropagation of errors. The foundations of this unique mathematical approach were laid by Paul Werbos in his groundbreaking 1974 Harvard doctoral thesis, whose full text is published in [Werbos 1994]. The algorithm is one of the most used by scientists, engineers and researchers involved in neural networks. A conceptual illustration of a three-layer Perceptron ANN is shown below:

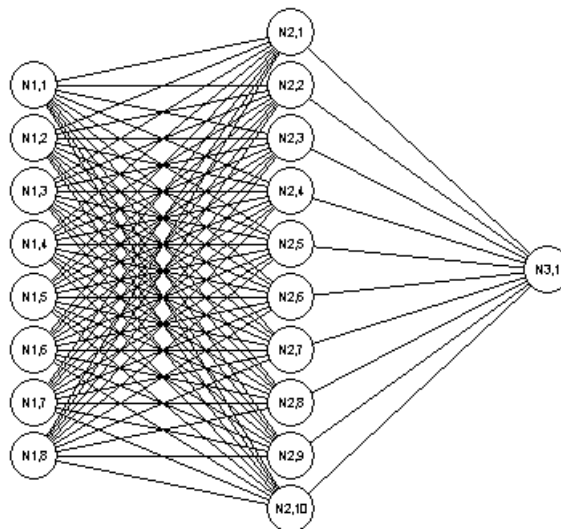


Figure 4.1: Conceptual Illustration of 3-Layer Perceptron ANN with Topology 8-10-1

In the example shown, it is a 3-layer Perceptron ANN with 8 Input Units, 10 Hidden Units, and 1 Output Unit. For simplicity, the topology, or number of units in each layer, is often given in the format X-Y-Z, where X represents the number of input units, Y represents the number of hidden units, and Z represents the number of output units.

For the experiments performed in this dissertation, a Perceptron ANN was constructed with 52 input units, 1 output unit, and a variable number of hidden units. To avoid non-linear modification of the input values (the features themselves), identical linear activation functions were used for all input neurons. Since a two-class delineation is required (negative or positive for variant lymphocytes), a single output neuron was used with bipolar targets. The chosen target designations will be -1 for samples negative for Variant Lymphocytes, and +1 for samples positive for Variant Lymphocytes. Again, to avoid making any assumptions about the output values produced, a linear activation function was initially chosen for the single output neuron. The backpropagation learning rule with a fixed learning rate was chosen to train the classifier.

The next issue tackled is in determining the need for hidden units. Initially, it will be assumed that acceptable classification accuracy for this problem can be found while using zero hidden units – making this neural network roughly equivalent to a linear classifier.

4.5. Data Partitioning

Out of the total 300 files used in this study, 150 (75 Variant Lymph Positive and 75 Variant Lymph Negative) hematology data files or 50% were selected randomly and used initially in a training phase to ascertain the reliability of the eigensystem-based feature space sample classification process. The remaining 150 data files or 50% were then used in the testing phase. The 150 files selected for training are from patients 1 to 150, and

patients 151 to 300 were used subsequently in the testing phase in order to validate the classifier's ability to perform well on an inter-patient level.

4.6. Artificial Neural Network Design Considerations

4.6.1. Type of Neural Network

Although the Perceptron was chosen for this classification problem using the 52-dimensional parameter space introduced in the previous section, it is important to emphasize that the choice of classification method is not crucial for this study, since the feature space has been demonstrated to have good discriminative power. Consequently; any machine learning framework is expected to achieve good classification results.

4.6.2. Learning Rule

Because the learning and weight update method "backpropagation of errors" has been shown mathematically to have sufficient ability, when coupled with multi-layer ANNs, to learn to find optimal decision functions by employing the method of gradient descent, it was decided to use this method for learning with the chosen Perceptron ANN.

4.6.3. Targets Used For Supervised Learning

Since a two-class delineation is required (negative or positive), a single output neuron was used with bipolar targets. Bipolar targets can be more advantageous than binary targets (0 and 1) because sometimes the value 0 can cause numerical overflow problems with floating point division operations that are common in gradient descent calculations; these overflow problems can lead to instability and/or premature stoppage in the training procedure by producing very large numbers. Bipolar targets (-1 and 1) force the targets to be equidistant and sufficiently further away from the zero value.

4.6.4. Type and Number of Input Units

The number of input units is fixed at 52, which is identical to the dimensionality of the input feature space. However, it may be determined that some inputs can be removed from the feature space if they do not contribute significantly to the ANN performance. In addition, to avoid nonlinear modification of the input values (the features themselves), no activation functions are used for the input neurons.

4.6.5. Activation Functions

Although the classification ability of an ANN is not strongly tied to the choice of activation functions used, they can have a noticeable effect on the quality and effectiveness of pattern learning [Fausett 1994] [Duda et al 2000].

4.6.6. Training Procedure

Typically, upon training artificial neural networks with supervised learning patterns, the issue of over-training (also termed over-fitting) is always a concern. The issue is that an ANN, if left to train for too long, can actually “memorize” all the patterns in the training set, and the classifier that is created will be of little value if it is used to classify “unseen” patterns. To assure that a classifier is not over-trained, and still will respond correctly to unseen patterns, cross-validation techniques must be employed during the neural network training/testing process. Once the optimal artificial neural network topology has been found, final training and testing of this network will, of course, always be done while using cross validation, to assure that the ANN network that is created maintains sufficient ability to correctly classify unseen data patterns [Weigend 1994] [Tetko et al 1995].

4.6.7. Learning Rates

A fixed, constant learning rate is normally sufficient for effective learning. It is important; however, to assure that the learning rate is large enough so as not to slow the learning process unnecessarily, but not be too large so as to cause oscillations or instability due to amplification of noise during the gradient descent optimization procedure. There is no exact criterion to determine the optimal learning rate for a particular ANN classification problem; it usually must be determined by “trial and error” procedures [Fausett 1994] [Swingler 1996].

4.6.8. Number of Hidden Units

Determination of this topological parameter can be critical to achieving a satisfactory ANN testing performance. Having no hidden units and linear activation functions on input and output units will generate a network that is equivalent to a linear classifier. Adding hidden units allows the network to approximate a more complex decision boundary. If the patterns are well-separated or linearly separable, then few hidden units are needed; however, if the contrary is true then larger numbers of hidden units will be necessary to obtain good classification results [Duda et al 2001].

4.7. Strategies to Find Optimal ANN Topology

For the task of deciding upon a network topology that can give acceptable accuracy for the data used in this study, some experiments were done with varying topologies. The following ANN parameters were varied in different experimental trials:

1. Activation Functions (Linear versus Sigmoidal or Tansig)
2. Number of Hidden Layers
3. Number of Hidden Units

4. Number of Input Features (All features versus smaller subsets)

4.7.1. Perceptron ANN Experiment #1

Initially, to avoid making assumptions about the feasibility of this classification problem, the most basic Perceptron ANN configuration is used. For exploratory purposes, it is often desired to see whether an input feature space (without the help of hidden units) can provide linear separability of the two pattern classes. This type of exploratory analysis is initially done without any cross-validation so that a reasonable assessment as to the potential or “best-case” separability that the data classes inherently contain; this is done intentionally knowing that some over-fitting of the ANN will likely occur.

For this experiment, the following Perceptron topology and parameters are used:

Table 4.2: Perceptron ANN Configuration Variables for Training Experiment #1

Number of Input Units	Number Of Hidden Units	Activation Functions Used	Learning Rate	Total Training Time (s)	Cross Validation Used (Y/N)
52	0	Linear	0.001	180	No

Using the entire 300 datasets, this network was trained for 3 minutes (roughly 4000 iterations per trial), 50 separate times so that an average training error result could be obtained. It was done with no cross-validation merely to show that even in the over-trained condition, with linear activation functions for both input and output neurons, that this ANN could not achieve 100% accuracy in a reasonable training time. This would provide support for the assertion that the two classes (Negative for variant lymphocytes and Positive for variant lymphocytes) are not linearly separable. The results for accuracy, sensitivity, and specificity for the 50 separate trials are given in Figure 4.2.

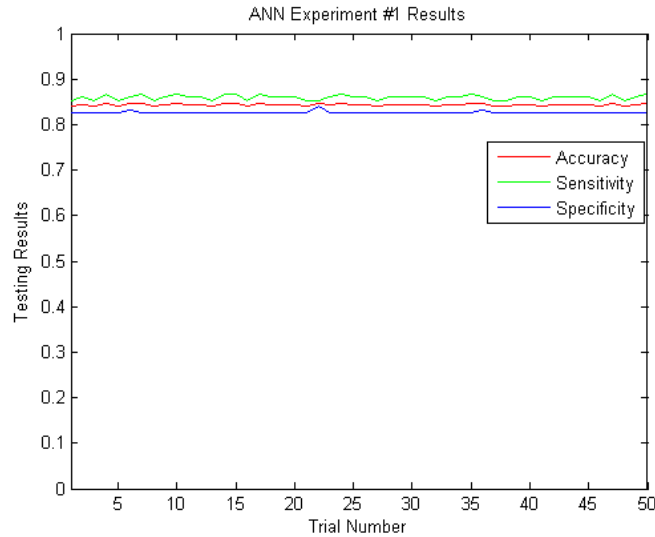


Figure 4.2: Accuracy, Sensitivity, and Specificity Results for 50 Training Trials – Experiment #1

Throughout the 50 trials, the best accuracy result obtained was for trial #22. The confusion matrix and performance statistics for this trial are given in Table 4.3.

Table 4.3: Confusion Matrix and Performance Statistics for Best Trial of Experiment #1

TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
128	126	24	22	84.67	85.33	84.00

Basically, the results did not change appreciably during the different training trials, which implied that there was little dependence of the testing result on the random initialization of weights (which was done for each trial automatically). Judging by the best accuracy result of only 84.7%, the 2 dataset classes (Negative for Variant Lymphocytes, and Positive for Variant Lymphocytes) are likely to be not linearly separable. Thus, it is fairly certain that a Perceptron with no hidden units and purely linear activation functions will not be able to provide perfect separation. Thus, to achieve optimal separation of these classes, the non-linear aspects of artificial neural networks (hidden units and

alternative activation function types) must be incorporated into the classification problem. Thus, the following trials introduce the usage of alternative activation functions and hidden units so that improvements in classification performance may be realized.

The activation functions of each neuron can be specified to be different mathematical transfer functions. Certain functions, such as the sigmoid, can be used to introduce non-linearity into the classification, thus potentially, improving the accuracy of the classifier for classes that are not linearly separable.

4.7.2. Perceptron ANN Experiment #2

The activation functions of each neuron can be specified to be different mathematical transfer functions, other than purely linear. Certain functions, such as the sigmoid, can be used to introduce non-linearity into the classification and thus potentially improve the accuracy of the classifier for classes that are not linearly separable.

Before resorting to the use of hidden layers and hidden units to achieve separation of the two classes of data, the use of a sigmoid function in the output unit is attempted. The equation of the logistic sigmoid (logsig) function that was used is given as the following:

$$F(x) = (y_{\max} - y_{\min}) \frac{1}{1 + \exp(-a(x - b))} + y_{\min} \quad (4.1)$$

In Equation 4.1, the a and b parameters are real numbers usually in the range of 0.1 to 5. The logsig function is often applied in backpropagation networks since its slope is not constant, and therefore, it can provide an effective way of updating weights depending on the value of the gradient of the error. Additionally, the logsig has the advantage that its slope y' can be expressed in terms of its output as $y' = y(1 - y)$, which makes it computationally practical. The following parameters for the sigmoid were used:

Table 4.4: Sigmoid Activation Function Parameter Values

Parameter	Value
a	3.0
b	0
Ymin	-1.0
Ymax	1.0

Using the parameters in Table 4.4 as inputs to Equation 4.1 yields the following function:

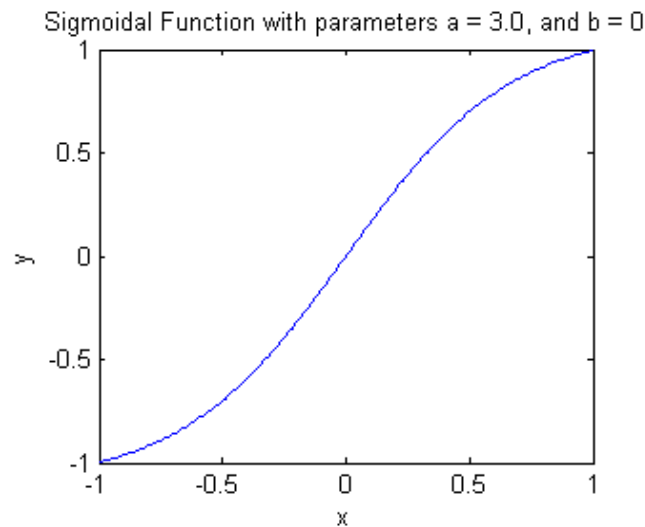


Figure 4.3: Graph of Sigmoid Function Using Parameters Defined in Table 4.1

For the ANN training experiment, the sigmoid, or logsig function specified in Figure 4.3 will be used as the activation function for all input units in the Perceptron. The classification performance of the network will of course be monitored to note the effect of this activation function. Again, using all 300 datasets for training and testing of the network, with no cross-validation techniques applied, the following Perceptron ANN topology and parameters were used:

Table 4.5: Perceptron ANN Configuration Variables for Training Experiment #2

Number of Input Units	Number Of Hidden Units	Activation Functions Used	Learning Rate	Total Training Time (s)	Cross Validation Used (Y/N)
52	0	Logsig	0.001	180	No

Again, this network was trained and tested 50 separate times. The performance was calculated on all 300 datasets each time. The 50 testing results are shown in Figure 4.4.

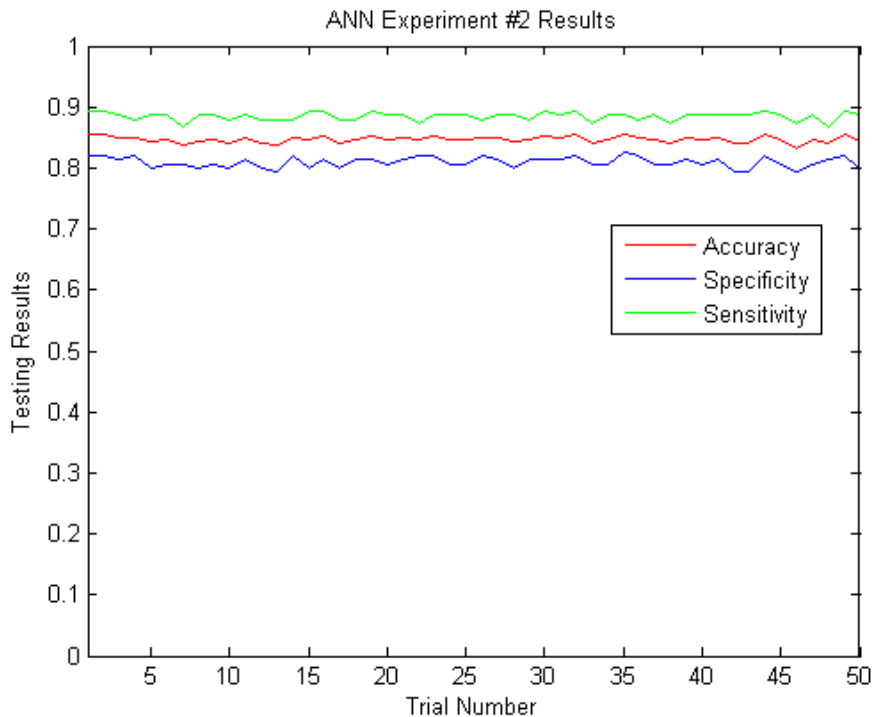


Figure 4.4: Accuracy, Sensitivity, and Specificity Results for 50 Training Trials – Experiment #2

Likewise, the best accuracy testing result (trial #35) is shown to assess the effect of adding the sigmoid activation functions. The confusion matrix and performance metrics of this trial are given in Table 4.6.

Table 4.6: Confusion Matrix and Performance Statistics for Best Trial of Experiment #2

TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
133	124	26	17	85.67	88.67	82.67

Again, it should be noted that the testing accuracy did not fluctuate between each trial, indicating that the testing results were repeatable, and not due to random chance. Comparing to the best results from Experiment #1, it was noted that Experiment #2 has slightly better overall accuracy (85.67% versus 84.67%), slightly better sensitivity (88.67% versus 85.33%) and slightly worse specificity (82.67% versus 84.0%). So, it is noted that using a sigmoid activation function on the output unit of the ANN had only a marginal, albeit improved effect on its accuracy.

4.7.3. Perceptron ANN Experiment #3

Since it seemed as though the ANN accuracy performance was not going to reach 100% through a single-layer network, even with sigmoidal activation functions, it seemed necessary to add a single hidden layer to the network to give it the ability to separate the pattern classes more effectively. Since the optimal number of hidden units is unknown, a simple test of different numbers of hidden units was done. A sweep of various numbers of hidden units from 20 to 120 was done and the average pattern error checked to see if an obvious trend would emerge. Since the number of input features is still fixed at 52, it did not seem logical to try a very small number of hidden units (e.g. 5 or 10). The number of hidden units was varied from roughly 40% of the inputs (20) to roughly 2.3 times the number of inputs (120). Again, this experiment was done without cross-validation, since the actual training results were not important; only the trend of training results relative to each other was sought.

Thus, the Perceptron ANN was again set up with the following topology and parameters:

Table 4.7: Perceptron ANN Configuration Variables for Training Experiment #3

Number of Input Units	Number Of Hidden Units	Activation Functions Used	Learning Rate	Total Training Time (s)	Cross Validation Used (Y/N)
52	Varied from 20-120	Logsig	0.001	180	No

To get a reliable result for each trial of different numbers of hidden units, 5 training trials were conducted for each number of hidden units, and an average of the pattern error for the 5 trials was computed and saved. The results of this analysis, average pattern error from 5 training trials, are shown as a function of number of hidden units in Figure 4.5.

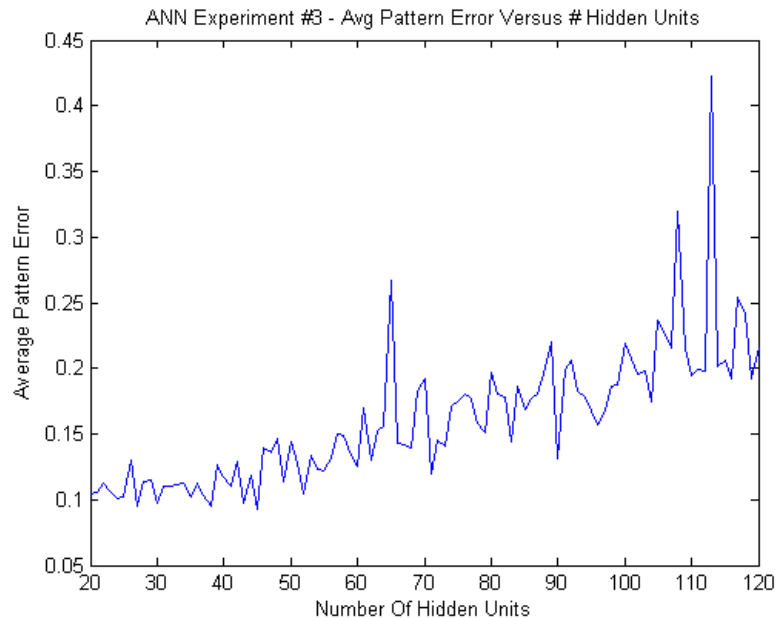


Figure 4.5: Average Training Pattern Error versus Number of Hidden Units

A general trend of increasing pattern error is seen as the number of hidden units increased beyond 50. In the range of 20 to 50 hidden units showed a fairly constant (and relatively low) average pattern error. It is possible that, in this training scenario, (without any cross-validation) 20-50 hidden units may be optimal for the best accuracy possible, since

there are no constraints imposed to avoid over-training the network. Is it therefore possible that a greater number of hidden units are necessary to construct a neural network with both the ability to perform well on training data as well as on unseen testing data. Thus, this experiment should be conducted in two trials (one with no cross-validation and one with cross-validation) and then the ultimate conclusion should be drawn. The technique of cross-validation is basically to divide the training set into mutually exclusive, but not equally-sized subsets, and for each subset, the classifier is trained on the union of all the other subsets. The average of the error rate of each subset is therefore an estimate of the error rate of the classifier. This type of validation is, of course, more demanding computationally, but useful when the most accurate estimate of a classifier's error rate is required [Kotsiantis 2007].

4.7.4. Perceptron ANN Experiment #4

Extending the coverage of Experiment #3 to address the effects of ANN over-training, the same experiment is now conducted using a technique of cross-validation. Cross validation will be implemented as follows: 50% of the available patterns will be used for training, 25% of the available patterns will be used for testing, and the remaining 25% will be used for cross-validation. The data for each of these sets will be selected in an equal-but-random fashion from the two sets of data (positive samples and negative samples), to assure that the number of positive samples and negative samples are always roughly equal, to avoid biases in the results. For the original dataset of 300 patterns (150 positive and 150 negative), the data will be partitioned as follows:

Table 4.8: Data Partitions for Training, Cross-Validation, and Testing

Data Subset	Number Of Files in Subset
Training	150 (75 Positive, 75 Negative)
Cross-Validation	75 (37 Positive, 38 Negative)
Testing	75 (38 Positive, 37 Negative)

Training was performed and finalized with early stopping (a cross-validation strategy) as a regularization procedure to avoid network memorization. The procedure was set as follows: every 3 iteration loops, the average square error on the cross-validation set is computed and compared to the previous 5 values computed thus far. If the last error is higher, the iterations are stopped, because this could represent an increasing error trend.

One may question the difference between the three data subsets (training, testing, and cross-validation). What is their distinction from one another? Why not just use training and testing sets? What is the purpose of having an additional set, the cross-validation set? The answer will require a short explanation. The training data will be iterated through the network again and again, over 1000's of iterations. It is expected that the ANN will eventually "memorize" each of the training patterns through the process of over-fitting. At this point, the ANN that is created may be relatively useless, because it possibly could not correctly recognize the "Variant Lymphocyte" trait or pattern in data samples presented to it that were not seen previously by the classifier during training. This is obviously a training pitfall that should be avoided [Weigand 1994]. That is where the testing data becomes important. It will not be involved in the training process in any

way; in fact, it will not be seen by the ANN. However, the cross-validation dataset differs from the testing dataset; it will not be trained upon either, but it used merely to determine if over training of the network has occurred. A common practice used in cross-validation is to allow the ANN to see all patterns of the training set N times. Then after it has been allowed to train for these N iterations, the cross-validation set will be evaluated by the ANN and an average pattern error for this subset of data will be computed. If this average pattern error for the cross-validation set continues to decrease (since this error will be computed after each N training iterations of the training data), the training process will continue for another N iterations [Tetko et al 1995]. It is important to realize that the cross-validation set is not trained upon by the ANN (since no weight updates will be caused by its usage), but it is merely used as a tool for monitoring whether sufficient (but not too much) training has been done for the ANN. Thus, employing this strategy will now give three sets of errors that must be monitored – training error on training data, testing error for testing data, and testing error for cross-validation data.

Thus, following similar parameters as Experiment #3, but with the important addition of the cross-validation technique to avoid over-training of the network, the following average pattern errors are seen as a function of number of hidden units:

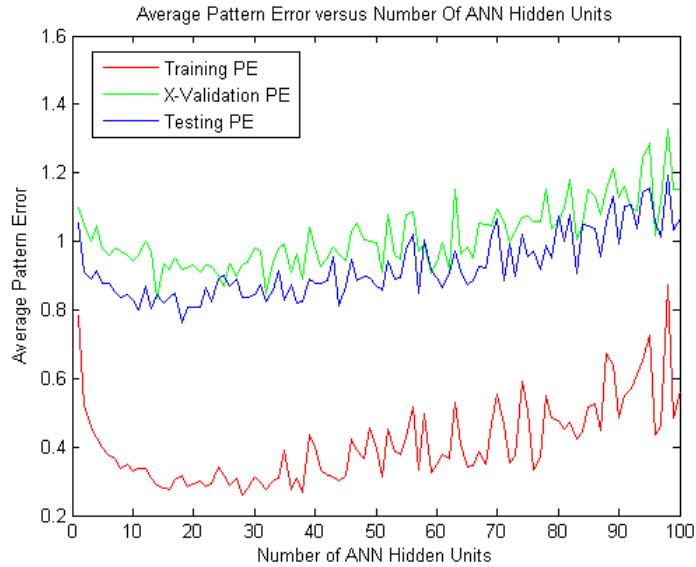


Figure 4.6: Average Pattern Error versus Number of Hidden Units with Cross-Validation

The average training pattern error is the lowest series (shown in red). The average training pattern error begins at a fairly high value when only 1 hidden units is used, then begins to trail off as the number of hidden units approaches 15. Then, there is a slightly upward trend for the average training pattern error as the number of hidden units increases beyond a certain point (close to 40). Between 15 and 40 hidden units, the average training pattern error is fairly constant.

This result may seem counter-intuitive because of the notion of over-fitting in the training process; we would expect this pattern error to decrease toward zero as the number of hidden units increases towards infinity. However, it is apparent that as the network complexity and degree of non-linearity increases (through the addition of more hidden units and their activation functions), the more prone it is to be unstable. The training error function of the network becomes more sensitive to perturbations, possibly due to the inherently difficult nature of solving highly-complex, and non-linear mathematical

functions for extremal values, which is what, in essence, a multi-layer Perceptron is attempting to do.

More important than average training error are the average pattern errors for the cross-validation and testing datasets. This error also clearly follows a discernible trend; their average pattern errors seem to be affected (although not as drastically as the training pattern errors) by the increase of number of hidden units. Also, it is clear from the data plots that the cross-validation set (shown in green) seems to always have a larger average pattern error than the testing set (shown in blue). So, to simplify this analysis (finding the optimal number of hidden units for this classification problem to have minimal error), we ignore the average pattern error for the training data (for the moment), and concentrate on the average pattern error for the other sets. For simplicity in viewing, an average of the pattern error for the two sets (cross-validation and testing) is shown versus the number of hidden units.

The results of this analysis are shown below:

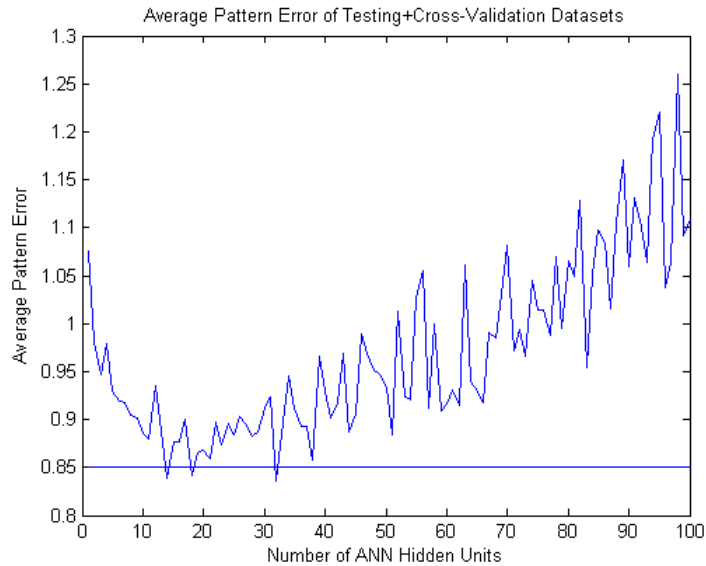


Figure 4.7: Average Pattern Error versus Number of ANN Hidden Units

Since the lowest average pattern error produced by untrained (unseen) data is ultimately what is desired from the final ANN design, this is the goal that is sought from this analysis. The lowest average pattern errors are seen at $N = 14$ (0.839), $N = 18$ (0.841), and $N = 32$ (0.837). This gives three potential trials for the optimal number of hidden units. Since the minimum pattern error was seen for $N = 32$ hidden units, that will be the first trial, followed by $N = 14$, and $N = 18$. The ANN topology demonstrating the best overall performance under the duress of repeated training and testing experiments (to assure that the good ANN performance is repeatable was not merely due to chance occurrence) will be chosen as final network topology.

4.7.5. Perceptron ANN Experiment #5

Using the same Perceptron inputs and parameters as those used in Experiment #4, and only allowing the number of hidden units to vary for ($N = 14$, 18, and 32), the following trials are performed.

4.7.5.1. Trial #1: N = 32 Hidden Units

Using N = 32 hidden units and the same Perceptron ANN setup with 52 inputs (linear activation function and no bias), sigmoid activation functions, and 1 output unit, the following performance was seen over 50 repeated training attempts:

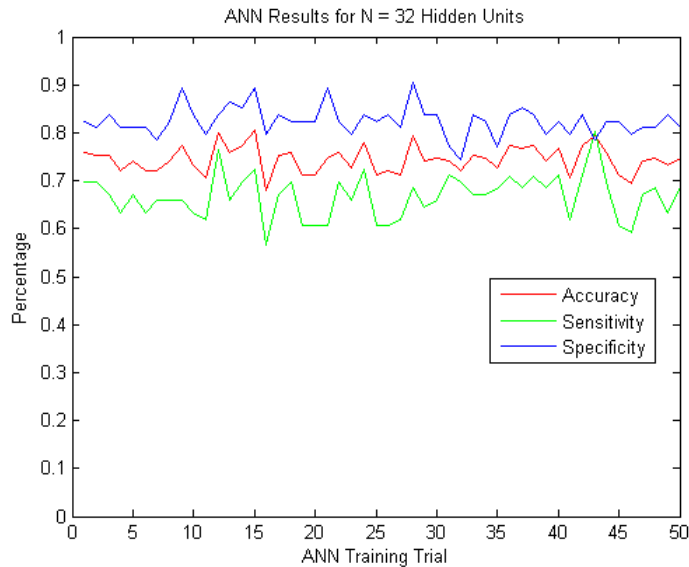


Figure 4.8: Accuracy, Sensitivity, and Specificity Results for 50 Training Trials – Experiment #5, Trial #1

The average performance metrics for the 50 training trials are shown in Table 4.9.

Table 4.9: Performance Result Ranges for Experiment #5, Trial #1

Accuracy Range	Sensitivity Range	Specificity Range
74.46 ± 2.79	66.68 ± 4.6	82.45 ± 3.11

Likewise, the best accuracy training result from the 50 trials is shown below:

Table 4.10: Confusion Matrix and Performance Statistics for Best Trial of Experiment #5, Trial #1

True Positive	True Negative	False Positive	False Negative	Accuracy	Sensitivity	Specificity
55	66	8	21	80.67%	72.37%	89.19%

4.7.5.2.Trial #2: N = 18 Hidden Units

The same procedure was repeated for this topology that contained 18 hidden units. The testing results for 50 separate trails are shown in Figure 4.9.

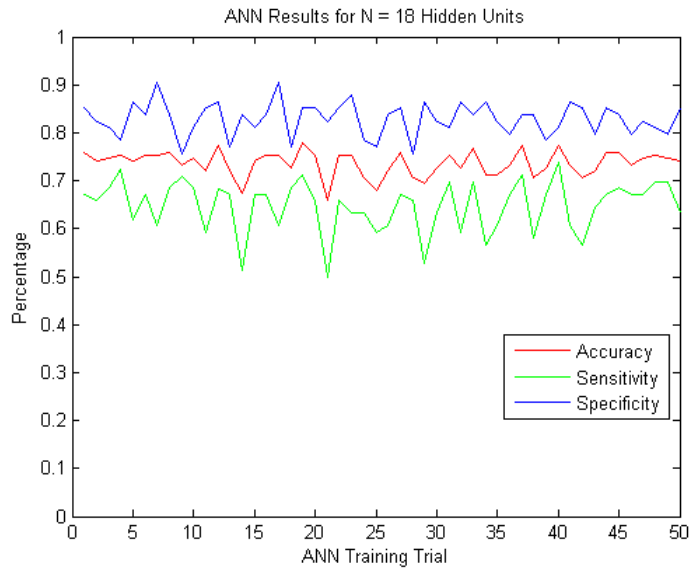


Figure 4.9: Accuracy, Sensitivity, and Specificity Results for 50 Training Trials – Experiment #5, Trial #2

The average performance metrics for the 50 training trials are also shown in Table 4.11.

Table 4.11: Performance Result Ranges for Experiment #5, Trial #2

Accuracy Range	Sensitivity Range	Specificity Range
73.65 ± 2.63	64.68 ± 5.4	82.86 ± 3.5

Likewise, the best accuracy training result from the 50 trials is shown below:

Table 4.12: Confusion Matrix and Performance Statistics for Best Trial of Experiment #5, Trial #2

True Positive	True Negative	False Positive	False Negative	Accuracy	Sensitivity	Specificity
54	63	11	22	78.0%	71.05%	85.14%

4.7.5.3.Trial #3: N = 14 Hidden Units

The same procedure was repeated for this topology that contained 14 hidden units. The testing results for the 50 separate trials are shown below in Figure 4.10.

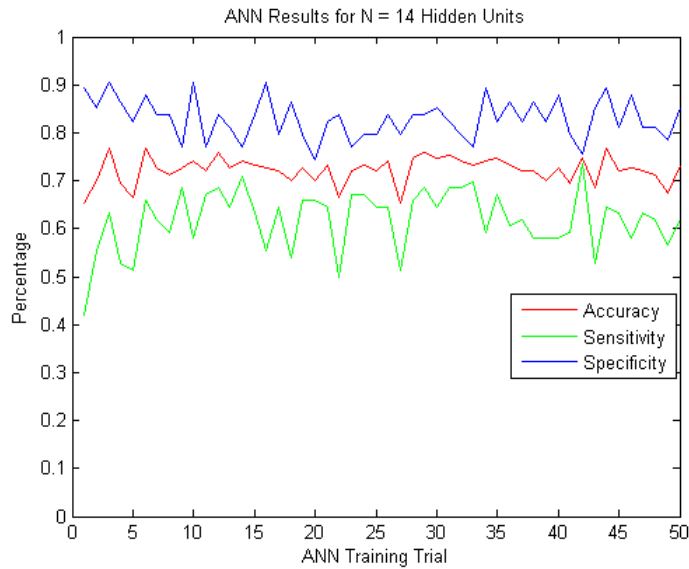


Figure 4.10: Accuracy, Sensitivity, and Specificity Results for 50 Training Trials – Experiment #5, Trial #3

The average performance metrics for the 50 training trials are shown in Table 4.13.:

Table 4.13: Performance Result Ranges for Experiment #5, Trial #3

Accuracy Range	Sensitivity Range	Specificity Range
72.24 ± 2.78	61.76 ± 6.28	83.0 ± 4.16

Likewise, the best accuracy training result from the 50 trials is shown below:

Table 4.14: Confusion Matrix and Performance Statistics for Best Trial of Experiment #5, Trial #3

True Positive	True Negative	False Positive	False Negative	Accuracy	Sensitivity	Specificity
49	66	8	27	76.67%	64.47%	89.19%

4.8. Discussion of Perceptron ANN Topology Experiments

It was shown that suitable values of the Perceptron parameter values (learning rate, activation function type and parameters, and number of hidden units) that would provide good classification performance for this problem could be found experimentally. The conclusions drawn from the five individual experiments can be summarized as follows:

Use of a linear classifier (Perceptron with no hidden units and linear activation functions on inputs and outputs) cannot give perfect separation of the two data classes, even when allowing “memorization” of all data to occur by foregoing any cross-validation activities. This conclusion was drawn in Experiment #1.

Sigmoid Activation Functions have some positive effects on the classification performance (small increase in accuracy). However, the addition of sigmoid activation functions alone was not sufficient to provide vast improvement in performance – these conclusions were drawn in Experiment #2.

Addition of a hidden layer is necessary to augment the performance capabilities of the ANN – this was seen in Experiments #3.

A suitable number of hidden units can be arrived upon by observing the average pattern error for multiple training results each attempted with increasing numbers of hidden units. Lower average pattern error indicates improved ANN performance. This was seen in experiments #4 and #5.

The optimal number of hidden units for this classification problem was chosen at the point when the average pattern error was minimized. Though 3 individual choices of N (number of hidden units) all gave very low values for average pattern error, it was noted that $N = 32$ hidden units gave the best set of average performance metrics for repeated

training scenarios. This statement is reinforced by observed average performance metrics for the 3 values of hidden units:

Table 4.15: Average Performance Metrics for 3 Topologies Tested in ANN Experiment #5

Number Of ANN Hidden Units	Average Accuracy Performance	Average Sensitivity Performance	Average Specificity Performance
32	74.46%	66.68%	82.45%
18	73.65%	64.68%	82.86%
14	72.24%	61.76%	83.0%

4.9. Finalized Perceptron ANN Topology and Best Performance

Assuming that a 3-layer (1 input layer, 1 hidden layer, 1 output layer) Perceptron will be used, and that the optimal ANN parameters (number of hidden units, activation functions, and learning rates) were already determined through the experiments detailed in Section 4.7, the following Perceptron ANN parameters will be used to demonstrate the following classification performance:

Table 4.16: Perceptron ANN Configuration Values of Optimal Network

Number of Input Units	52
Input Unit Activation Function	Linear
Number of Hidden Layers	1
Number of Hidden Units	32
Hidden Unit Activation Functions	Sigmoid
Activation Function Parameters	See Table 4.4
Number Of Output Units	1
Output Unit Activation Functions	Sigmoid
Activation Function Parameters	See Table 4.4
Global Learning Rate	0.001
Total Training Time	180 seconds
Cross-Validation Used For Training	Yes

It is noteworthy to mention that the Perceptron ANN with $N = 32$ hidden units and the other setup parameters mentioned in Table 4.16 demonstrated the best combination of accuracy, sensitivity, and specificity seen to date.

The best overall results shown from this ANN topology (which represents the optimal classifier for Variant Lymphocyte detection as determined through this series of experiments), with respect to the Coulter LH 750 Variant Lymphocyte classifier are shown in the following tables. The classification accuracy, sensitivity, and specificity was evaluated using the Cross-Validation and Testing datasets combined (total of 150 files). This was done because the training dataset (150 files) was essentially memorized by the Perceptron ANN and does not give an unbiased testing result. So, using the 150 files designated for testing and cross-validation only, the side-by-side comparison of performance between the two classifiers on exactly the same set of data is shown below:

LH750 Variant Lymphocyte Classifier Performance:

Table 4.17: Performance of LH750 Variant Lymph Suspect Flag on Testing + Cross-Validation Datasets

True Positive	True Negative	False Positive	False Negative	Accuracy	Sensitivity	Specificity
22	70	4	54	61.33%	28.94%	94.59%

Eigensystem-Feature Variant Lymphocyte Classifier Performance:

Table 4.18: Performance of Proposed Variant Lymph Suspect Flag on Testing + Cross-Validation Datasets

True Positive	True Negative	False Positive	False Negative	Accuracy	Sensitivity	Specificity
55	66	8	21	80.67%	72.37%	89.19%

It is fairly obvious that the current variant lymphocyte classifier implementation using the Perceptron ANN with 52 input Eigen-parameter features gives improved performance.

This is evident not only from overall accuracy (80.67% versus 61.33%), but also from overall sensitivity (72.37% versus 28.94%). The only drawback of this implementation is that its specificity was not as high as the Coulter LH 750 implementation (89.19% versus 94.59%). However, it is felt that the 5.4% increase in false positives is an acceptable tradeoff for the vastly improved sensitivity and accuracy of the implemented approach.

5. CLASSIFICATION OF VARIANT LYMPHOCYTES USING FEATURE EXTRACTION AND MACHINE LEARNING

5.1. Objectives of the Method

The main objective of this method was to demonstrate that the set of multi-dimensional eigensystem-based hematological population descriptors that was proposed by this dissertation is clearly more effective for correctly classifying samples containing medically-relevant levels of variant lymphocytes than are basic 1-dimensional population statistical features such as mean and standard deviation.

5.2. Data Collection

For this study, hematological data collected using a Beckman Coulter LH750 automated hematology analyzer. Since this study focused on the classification a specific type of WBC abnormality, the presence of Variant Lymphocytes, only the data from the WBC differential mode of operation of the LH750 analyzer was used in this study. Specifically, this amounted to data from 300 individual donors. The first 150 donors were randomly chosen because they were negative for medically-abnormal levels of Variant Lymphocytes, and the next 150 donors were randomly chosen because they were positive for medically-abnormal levels of Variant Lymphocytes.

5.3. Feature Extraction

As stated earlier, a set of 52 parameters was extracted from the 4 most significant WBC populations in each hematology sample, which are the Lymphocyte, Monocyte, Neutrophil, and Eosinophil populations.

Table 5.1: Donor Information and Data Used

Donor	Manual Variant Lymphocyte Percent	Is Donor Medically-Positive for Variant Lymphocytes
1	1.75	No
2	0.5	No
3	4.8	No
4	2	No
5	0.25	No
.....		
150	0.75	No
151	8.5	Yes
152	22.5	Yes
153	5	Yes
154	12	Yes
.....		
300	6	Yes

5.3.1. ANN Configuration and Training Procedure

For the problem at hand, a 52-32-1 Perceptron Artificial Neural Network topology was ultimately selected. The rationalization for selecting 32 hidden units and no other number of hidden units was made through interpretation of the results of the ANN topology experiments that had been reported in section 4.7, which indicated that an optimization of the number of hidden units for the problem at hand was possible. Even though many of the parameters used for the final network were experimentally determined through trial and error procedures, examination of the average pattern error

and classification accuracy performance upon completion of each trial provided the overall assurance whether specific parameter choices were more beneficial for the problem at hand. It was assumed that parameter choices (activation function types and parameters, learning rates, and number of hidden units) that provided the lowest average pattern errors during the testing phase were most likely to be optimal choices for the Variant Lymphocyte classification framework.

The input units were assigned linear activation functions, and the hidden and output units were assigned logsig functions, with the same parameters that were used in section 4.9.

The cross-validation strategy implemented in this case was identical to the one applied in section 4.9. Training was stopped once an increment in the cross-validation set for 5 consecutive times was detected. Under normal conditions, a training error can be 50% higher or lower than the previous one. By allowing the error no more than 5 consecutive times to increase, one is assuming that the error will keep increasing with a probability of $1 - (0.5)^5 = 1 - 0.03125 = 0.96875$, which is a high-confidence value. The learning rate was set in all ANNs to 0.001.

Because the minimum error obtained by an ANN depends on the starting condition, there is no way to know in advance which will be the best set of weights that will lead the network to a global minimum. The only way is to perform several trials, each starting with random weights, and to store the set of weights that yielded the best solution. This study opted to find the best solution for each ANN after a number of trials.

For each trial, the ANN was trained 50 times and the average performance results (average accuracy, average sensitivity, and average specificity) were shown as representations of the performance. This is done since a single training result can be

influenced heavily by the random starting values of the weights and biases. Thus, it is assumed that average performances values created with $N = 50$ trials are statistically significant. The intent of this approach with so many repeated trainings is to rule out the possibility that any differences in the results are just by chance (analogous to the statistical T-test).

Since cross-validation was performed, all training repetitions were stopped either by the cross-validation stop criterion or by reaching the time limit that was set. The procedure was applied consistently across all ANN topologies. In general, each ANN was allocated a time span of 3 minutes for training.

As it is known, cross-validation stops only when the validation error starts to increase, as shown in Figure 5.1. The number of iterations when the network begins to over-fit and loses its generalization ability is labeled in the figure as $N_{critical}$. But before reaching that point, the error on the testing set can either increase or decrease, although most of the times this error is expected to follow the trend of the cross-validation error, at least for the first few iterations. In this experiment, this tendency was observed in most trainings iterations, but not in all. Because the training algorithm has purposely no feedback from the testing error, limiting the search time is a good way to avoid high testing errors.

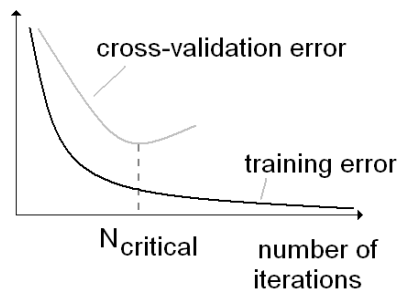


Figure 5.1: Illustration of the cross-validation criterion to stop training.

5.4. Testing Results

Since the object of this dissertation was to demonstrate that the 52-dimensional eigensystem population feature descriptor matrix could provide improved Variant Lymphocyte classification performance, the performance of the implemented approach was compared against a well-known Variant Lymphocyte classification algorithm which is provided as part of the Beckman Coulter LH750 Automated Hematology Analyzer. The classification performance of this analyzer's Variant Lymphocyte suspect flag on the 150 data files designated for testing by this study is shown in Table 5.2.

Table 5.2: LH750 Variant Lymphocyte Suspect Flag Performance

True Positive	True Negative	False Positive	False Negative	Accuracy	Sensitivity	Specificity
22	70	4	54	61.33%	28.94%	94.59%

Using the best individual testing result from the 50 trials of the Perceptron ANN topology showing the highest average accuracy performance (the Perceptron with topology 52-32-1) in the various experimental trials explained in Chapter 4, the performance using the same 150 testing data files is shown again in Figure 5.3.

Table 5.3: Eigensystem-Based Variant Lymphocyte Suspect Flag Performance

True Positive	True Negative	False Positive	False Negative	Accuracy	Sensitivity	Specificity
55	66	8	21	80.67%	72.37%	89.19%

5.5. Discussion of Classification Performance Improvements Obtained

It is notable to mention that the accuracy performance of the implemented Variant Lymphocyte detection method has increased from 61.3% to 80.67%, or by approximately a factor of 1.31, or 31%. Likewise, the sensitivity of the implemented approach has

increased from 28.94% to 72.37%, or by approximately a factor of 2.41. The downfall is the decrease in specificity by a factor of 0.94. This, however, seems to be an acceptable tradeoff, since accuracy and sensitivity have increased dramatically with only a modest decrease in specificity.

5.6. Samples Classified by Implemented Approach

Some patterns that were incorrectly detected by the LH750 Variant Lymphocyte suspect flagging method are now discussed. Many of these cases have been correctly detected by the implemented Variant Lymphocyte detection approach as discussed next.

5.6.1. Pattern #1: Variant Lymph Positive Sample Correctly Detected

Below is an example of a file that was not detected as having a medically-abnormal level of variant lymphocytes by the LH750 Variant Lymphocyte suspect flag, thus it is treated as a False Negative.

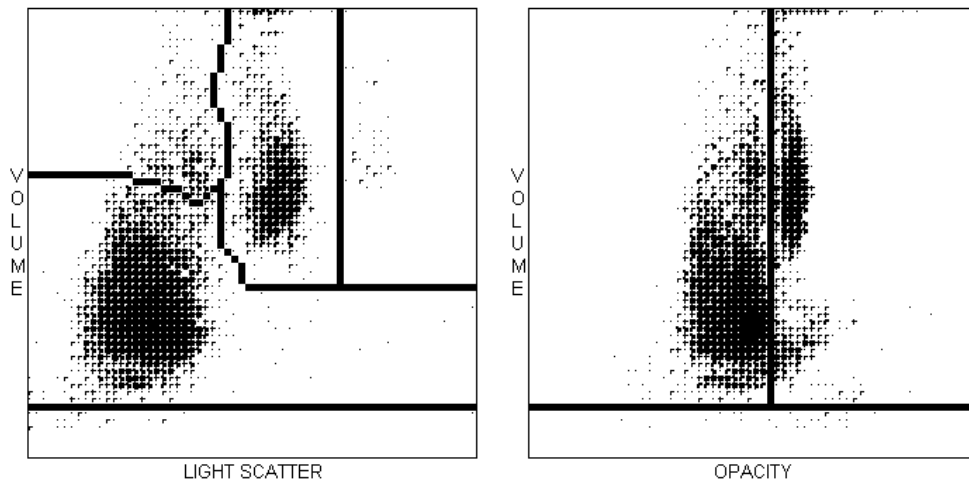


Figure 5.2: Variant Lymph Positive Sample – Discussion Pattern #1

The manual differential is also given for this donor, so that the seriousness of this sample may be understood. In particular, this case is serious because of the extremely high percentage of variant lymphocytes seen, which was reported as 35.25% according to the

manual differential report. This case ideally should be flagged as containing Variant Lymphocytes, and to not do so may be medically dangerous. This sample also has a very elongated lymphocyte population in the volume and opacity directions; this is a likely reason why the current classification method detected it as a true positive sample.

Table 5.4: Manual WBC Differential for Discussion Pattern #1

Lymph	Mono	Neutro	Eo	Baso	Blast	Immature Gran	Immature Band	Variant Lymph	Plasma Cell
51.25	3.0	9.5	0.75	0	0	0	0.25	35.25	0

5.6.2. Pattern #2: Variant Lymphocyte Negative Sample Incorrectly Detected

The following pattern is very difficult to detect correctly using the implemented approach. This sample exhibits a very large lymphocyte percent, and a fairly abnormal pattern appearance, yet, according to the manual differential, it does not contain a medically-abnormal level of Variant Lymphocytes:

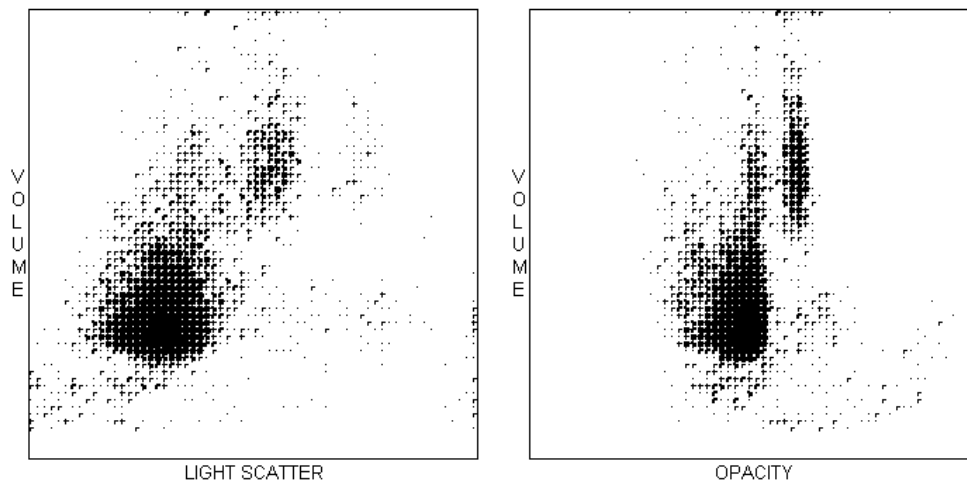


Figure 5.3: Variant Lymph Negative Sample – Discussion Pattern #2

The manual differential is shown for this donor as follows:

Table 5.5: Manual WBC Differential for Discussion Pattern #2

Lymph	Mono	Neutro	Eo	Baso	Blast	Immature Gran	Immature Band	Variant Lymph	Plasma Cell
91.75	1.0	4.75	0.75	0	0	0	0	1.75	0

This pattern is abnormal in appearance, but the manual differential states that a very low level of variant lymphocytes exists (less than the critical 5% value determined by medical doctors); thus the correct classification would be to declare that this sample is negative for variant lymphocytes. However, this sample is a False Positive for the current implementation. On the other hand, the Coulter LH750 variant lymphocyte suspect flag gives the correct classification for this sample.

5.6.3. Pattern #3: Variant Lymph Positive Sample Correctly Detected

A difficult pattern to detect is the type of sample that is clinically positive for variant lymphocytes but for which the manual differential states that less than 10 percent of variant lymphocytes are present; thus these are samples with between 5 and 10 percent Variant Lymphocytes. This problem was stated in several hematological publications as a major limitation of the Variant Lymphocyte detection algorithms on several commercially manufactured automated hematology analyzers [Aulesa et al 2003] [Aulesa et al 2004] [Hoffmann and Hoedemakers 2004]. This problem is usually attributed to an apparent lack of sensitivity that most Variant Lymphocyte detection methods seem to possess. For this reason, the implemented approach made an effort to greatly improve the sensitivity of detection of samples in the 5-10% Variant Lymphocyte range, while at the same time, not increasing appreciably the level of False Positives.

This is clearly not an easy task. An example of this type of Variant Lymphocyte pattern is shown below:

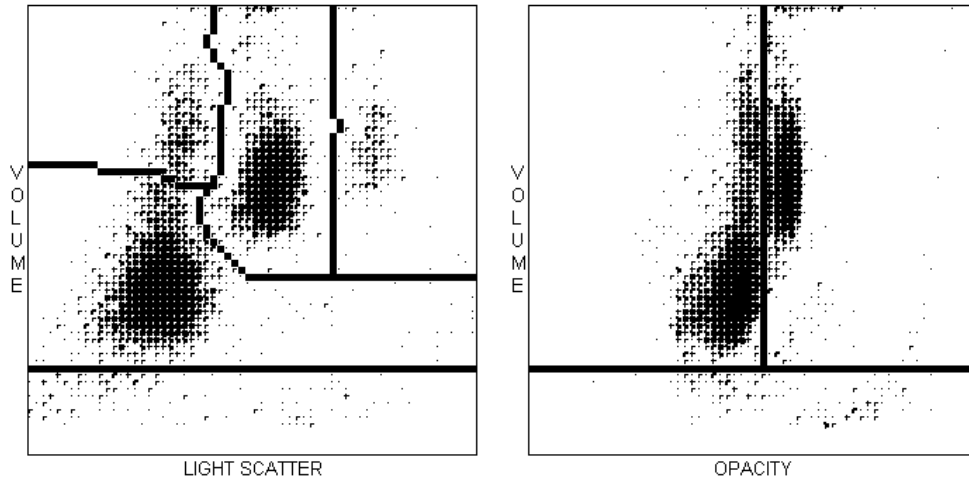


Figure 5.4: Variant Lymph Positive Sample – Discussion Pattern #3

The WBC manual differential report is given for this donor:

Table 5.6: Manual WBC Differential for Discussion Pattern #3

Lymph	Mono	Neutro	Eo	Baso	Blast	Immature Gran	Immature Band	Variant Lymph	Plasma Cell
60.0	4.5	26.5	2.5	0	0	0	0	6.5	0

It should be noted that only 6.5% variant lymphocytes are present, and the pattern does not appear to be extremely abnormal, thus, from a pattern detection standpoint, this is an example of a pattern for which detection of variant lymphocytes can be challenging. Since the proposed method takes into consideration not only the Lymph and Mono percents, but also the shape and orientation statistics (due to use of the eigenvectors and eigenvalues) of each population; for this reason, the classifier was able to detect an abnormal shape signature (perhaps due to the overlap between the Lymph and Mono populations). Thus, for this sample, the proposed method was able to detect this sample

as having an abnormal level of variant lymphocytes while the LH750 Variant Lymph detection approach was not.

For the reasons posed in these examples, it is felt that the implementation of the Variant Lymphocyte detection using the 52-parameter eigensystem descriptor matrix offers a better accuracy, and superior sensitivity than the existing Variant Lymphocyte detection method used by the LH750 automated hematology analyzer.

Since the object of this dissertation was to demonstrate that the 52-dimensional eigensystem population feature descriptor matrix could provide improved Variant Lymphocyte classification performance, the performance of the implemented approach was compared against a well-known Variant Lymphocyte classification algorithm, which is provided as part of the Beckman Coulter LH750 Automated Hematology Analyzer product.

6. CONCLUSIONS

The main intention of this dissertation was to present the proposed eigensystem-based cluster shape descriptors and show their applicability to a well-known pattern classification problem in clinical hematology which was also fully addressed from a medical application perspective and related mathematical framework.

The multi-dimensional set of shape descriptor features developed was demonstrated, through the application of machine learning frameworks such as Artificial Neural Networks, to be capable of detecting a class of hematological patterns more effectively and accurately than existing features, attesting to the attractiveness of this novel approach.

Chapter 1 of this dissertation gives an overview of recent applications of Artificial Intelligence to problem solving in several areas of biomedical research. A foundation for understanding the primary issues and goals in clinical hematology was provided in Chapter 2, which also set the stage for understanding the significance of current limitations and difficulties faced in the automated detection of clinical abnormalities in the hematology field. This chapter was crucial for understanding the basics of the methods later developed in this dissertation.

With a focus on detecting hematological patterns corresponding to blood samples containing medically-abnormal levels of Variant Lymphocytes, an in-depth review of the some of the various hematological pattern manifestations that are commonly encountered with respect to this disorder was given. Also, a description of the commonly-used feature extraction methods for describing hematological samples containing variant lymphocytes

was presented in Chapter 3. Some limitations of current features for the purposes of pattern classification were then expressed. A novel set of feature descriptors was then proposed.

In pursuing practical applications in the area of automated hematology analyses, Chapter 4 described the design and implementation of a method aimed at detecting hematological patterns that exhibit medically-abnormal levels of Variant Lymphocytes. The data was provided by Beckman Coulter Corporation and involved 300 patients who had their blood analyzed by a LH750 automated hematology analyzer. The approach implemented consisted of extracting a 52-dimensional feature space that is descriptive of the physical spread and orientation of each WBC population cluster in three-dimensional space as features to identify medically abnormal hematology patterns.

It was found that the designed Perceptron ANN classifier created using the eigen-system statistical descriptors as inputs yielded 80.67% accuracy, 72.37% sensitivity, and 89.19% specificity on the data chosen for this research study. Compared to the Beckman Coulter LH750 Variant Lymphocyte detection method, which produced 61.3% accuracy, 28.9% sensitivity, and 94.6% specificity, the classification performance of the proposed feature space and machine learning methods implemented in this dissertation represent a notable improvement.

Within the experimentally-determined optimal range of 15 to 40 hidden units, increasing the number of hidden neurons did not improve the results. This observation was made after a large number of training iterations, all stopped with cross-validation. Possible explanations for this can be as follows: 1) starting points for the iterations are always different and 2) local minima are always possible. Since the backpropagation algorithm

used in training the networks is always started from random solutions to avoid local minima, it is perfectly possible to achieve higher testing and even training errors if more hidden units are added. To reach this conclusion, up to ten repetitions were done and averaged in each topology. More repetitions could have been performed for each topology, however, with increasing data size and number of neurons, the time needed for the optimization increases so that the test becomes impractical.

The algorithm developed here might be capable of making a substantial contribution to the diagnostic gain of detected hematological abnormalities using pattern-based frameworks.

In Chapter 5, a discussion of the classification results introduced in Chapter 4 was made. Specific case files that were not detected correctly by the Beckman Coulter Variant Lymphocyte detection method employed on the LH750 hematology instrument were reviewed. For many of these cases, the eigensystem-based feature classification framework proposed in this dissertation was able to detect many of these cases correctly, attesting to the contribution of this feature space as a basis for abnormal pattern detection. The method was based on establishing a descriptor matrix for a hematological data pattern which proved to be advantageous in terms of its detection abilities. The algorithm for computing the matrix is straightforward and involves statistical operations on features derived from population data point arrays. The attractiveness of this feature matrix is that it is a shift-and-rotation-invariant representation that can be used to represent and compare hematological patterns with different shape, orientation, and appearance characteristics by using this single set of descriptors.

BIBLIOGRAPHY

- Aulesa C, Pastor I, Narajo D, and Galimany R, "Application of Receiver Operating Characteristics Curve (ROC) Analysis When Definitive and Suspect Morphologic Flags Appear in the New Coulter LH 750 Analyzer", *Laboratory Hematology* 10(1), pp. 14-23, March 2004.
- Baraldi A, Blonda P, A Survey of Fuzzy Clustering Algorithms for Pattern Recognition - Part II, *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, Dec 1999, Volume 29, Issue 6, ISSN: 1083-4419, pp. 786-801.
- Bessman JD, *Automated Blood Counts and Differentials*, Baltimore and London: The Johns Hopkins University Press, 1986.
- Caldwell CW, Lacombe F, *Evaluation of Peripheral Blood Lymphocytosis*, Academic Information Systems, 2000.
- Ding C, Zhou D, He X, Zha H, R1-PCA: Rotational Invariant L1-Norm Principal Component Analysis for Robust Subspace Factorization, *Proceedings of the 23rd International Conference on Machine Learning*, 2006. ISBN 1-59593-383-2.
- Dimitrov V, Korotkich V, *Fuzzy Logic: A Framework for the New Millennium*, Springer, 2002, ISBN 3790814253, 9783790814255.
- Drouet M, Lees O, *Clinical Applications of Flow Cytometry in Hematology and Immunology*, *Biol Cell*, 1993, Volume 78, pp. 73-78.
- Duda R, Hart P, Stork D, *Pattern Classification*, NY: John Wiley & Sons, 2nd Edition, 2001.
- Fausett L, *Fundamentals of Neural Networks*, NJ: Prentice Hall, 1994.
- Fogel DB, *Evolutionary Computation: Towards a New Philosophy of Machine Intelligence*. New York: IEEE Press 2000, pp. 140.
- Fogel LJ, Owens AJ, Walsh MJ, *Artificial Intelligence through Simulated Evolution*, Wiley, New York, 1996.
- Fraser A, *Simulation of Genetic Systems by Automatic Digital Computers. I. Introduction*, *Aust. J. Biol. Sci.*, 1957, Volume 10, pp. 484-491.
- Fraser A, Burnell D, *Computer Models in Genetics*. New York: McGraw-Hill, 1970.
- Gonzalez R, Woods R, *Digital Image Processing*, New Jersey: Prentice Hall, 2nd Edition, 2002.

Hoffmann J, Hoedemakers R, Diagnostic Performance of the Variant Lymphocyte Flag of the Abbott Cell-Dyn 4000 Hematology Analyzer, *Clinical and Laboratory Hematology*, February 2004, Volume 26, Number 1, pp. 9-13.

Horner MJ, Ries L, Krapcho M, Neyman N, *Seer Cancer Statistics Review: 1975-2006, Surveillance Epidemiology and End Results (SEER)*, Bethesda, Md., National Cancer Institute, November 2009.

Jackson JE, *A User's Guide to Principal Components*, NY: John Wiley & Sons, 1991.

Jolliffe IT, *Principal Component Analysis*, NY: Springer-Verlag, 1986.

Kim HK, Kim JD, Region-Based Shape Descriptor Invariant to Rotation, Scale, and Translation, *Signal Processing: Image Communication*, Volume 16, Number 1-2, 2000, pp. 87-93.

Klir GJ, Yuan B, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Prentice Hall PTR, 1995, ISBN 0131011715, 9780131011717.

Koepke J, Van Assendelft O, Brindza L, Davis B, Fernandes B, Gewirtz A, Rabinovitch A, Reference Leukocyte (WBC) Differential Count (Proportional) and Evaluation of Instrumental Methods, 2nd Edition, *Clinical and Laboratory Standards Institute*, January 2007, Volume 27, Number 4.

Kohavi R, Provost F, Glossary of Terms, *Machine Learning*, 30, 1998, pp. 271-274.

Kohonen T, *Self-organizing Maps*, 3rd Edition, Springer 2001, ISBN 3540679219, 9783540679219

Kothari R, Cualing H, Balachander T, Neural Network Analysis of Flow Cytometry Immunophenotype Data, Context-based Automated Detection of Epileptogenic Sharp Transients in the EEG: Elimination of False Positives, Aug. 1996, Volume 43, No. 8, pp. 803-810.

Kotsiantis SB, *Supervised Machine Learning: A Review of Classification Techniques*, Informatica, 2007, Volume 31, pp. 249-268.

Kriegel H, Kroger P, Schubert E, Zimek A, A General Framework for Increasing the Robustness of PCA-Based Correlation Clustering Algorithms" *Proceedings of the 20th International Conference on Scientific and Statistical Database Management (SSDBM)*, Hong Kong, China, 2008.

Kwak, N, "Principal Component Analysis Based on L1-Norm Minimization", IEEE Transactions on Pattern and Machine Intelligence, September 2008, Volume 30, Number 9, pp. 1672-1680.

Leukemia and Lymphoma Society. Leukemia, Lymphoma, Myeloma Facts 2010-2011. The Leukemia and Lymphoma Society, 2011.

Lorenson W, Cline H, Marching Cubes: A High Performance 3D Surface Reconstruction Algorithm, Computer Graphics, 1987, Volume 21, Number 4, pp. 163-169.

Marcum J, Statistical Theory of Target Detection by Pulsed Radar, IEEE Trans. Info. Thry., Apr. 1960.

Minsky ML, Papert SA, Perceptrons (Cambridge, MA: MIT Press), 1969.

Reddick WE, Mulhern RK, Elkin TD, Glass JO, Merchant TE, Langston JW, A Hybrid Neural Network Analysis of Subtle Brain Volume Differences in Children Surviving Brain Tumors, Magnetic Resonance Imaging, Elsevier Science, New York, NY, ISSN 0730-725X, 1998, Volume 16, No 4, pp. 413-421.

Rich E, Knight K, Artificial Intelligence, McGraw-Hill, New York, 1991, pp. 105-130.

Ripley BD, Pattern Recognition and Neural Networks, 8th Edition, Cambridge University Press, 1996, ISBN 0521460867, 9780521460866.

Rodak BF, Fritsma GA, Doig K, Hematology Clinical Principles and Applications, 4th Edition, St. Louis, Missouri, Saunders Elsevier, 2007.

Rosenblatt F, The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Cornell Aeronautical Laboratory, Psychological Review, 1958, Volume 65, No. 6, pp. 386-408.

Sarle WS, Stopped Training and Other Remedies for Overfitting, Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics, 1995, pp. 352-360, Available at: <ftp://ftp.sas.com/pub/neural/inter95.ps.Z>

Specht DF, Probabilistic Neural Networks and the Polynomial ADALINE as Complementary Techniques for Classification, IEEE Transactions on Neural Networks, Mar 1990, Volume 1, Issue 1, ISSN: 1045-9227, pp. 111-121.

Swingler K, Applying Neural Networks: A Practical Guide, London: Academic Press, 1996.

Tetko IV, Livingstone DJ, Luik AI, Neural Network Studies 1, Comparison of Overfitting and Overtraining, J. Chem. Info. Comp. Sci., 35, 1995, pp. 826-833.

Tilbury J, Eetvelt P, Garibaldi J, Curnow J, Ifeachor E, Receiver Operating Characteristic Analysis for Intelligent Medical Systems - A New Approach for Finding Confidence Intervals, IEEE Transactions on Biomedical Engineering, 2000, Volume 47, No.7, pp. 952-963.

Turgeon ML, Clinical Hematology: Theory and Practice, Lippincott, Williams & Wilkins, Baltimore, Md, 4th Edition, 2005.

Van der Meer W, Van Gelder W, De Keijzer R, Willems H, The Divergent Morphological Classification of Variant Lymphocytes in Blood Smears, Journal of Clinical Pathology, 2007, Volume 60, Number 7, pp. 838-839.

Vapnik VN, The Nature of Statistical Learning Theory, New York: Springer-Verlag, 1995.

Weigend A, On Overfitting and the Effective Number of Hidden Units, Proceedings of the 1993 Connectionist Models Summer School, 1994, pp. 335-342.

Werbos PJ, The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting, Series on Adaptive and Learning Systems for Signal Processing, Communications and Control, Wiley-Interscience, John Wiley & Sons, Inc., New York, NY 10158-0012, Feb 1994, ISBN 0-471-59897-6.

Widrow B, Lehr MA, 30 years of Adaptive Neural Networks: Perceptron, MADALINE, and Backpropagation, Proceedings of the IEEE, Sep 1990, Volume 78, Issue 9, ISSN: 0018-9219, pp. 1415-1442.

Zini G, D'Onofrio G, Neural Networks in Hematopoietic Malignancies, Clinica Chimica Acta, July 2003, Volume 333, Number 2, pp. 195-210.

Zadeh LA, Fuzzy Sets, Information and Control 8, 338-353, 1965.

Zong N, Adjouadi M, Ayala M, Artificial Neural Networks Approaches for Multidimensional Classification of Acute Lymphoblastic Leukemia Gene Expression Samples, WSEAS Transactions on Information Science and Applications, Volume 2 (8), August 2005, pp. 1071-1078.

Zong M, Adjouadi M, Ayala M, Optimizing the Classification of Acute Lymphoblastic Leukemia and Acute Myeloid Leukemia Samples using Artificial Neural Networks, Biomedical Sciences Instrumentation, Volume 42, pp. 261-266. Also presented at the Rocky Mountain Bioengineering Symposium, Terre Haute, Indiana, April 7-9, 2006.

VITA

MARK ALEXANDER ROSSMAN

EDUCATION

- | | |
|----------------|--|
| 1994 - 1999 | B.S. in Electrical Engineering, Florida International University, Miami, Florida |
| 1999 - 2003 | M.S in Computer Engineering, Florida International University, Miami, Florida |
| 2004 - present | Senior Software Engineer
Beckman Coulter Corporation, Miami, FL. |

PUBLICATIONS AND PRESENTATIONS

1. Rossman M, Candocia, F, and Adjouadi, M, Jayakar P, and Yaylali, I, Application of Affine Transformations for the Co-registration of SPECT Images, Proceedings of the Fourth IASTED International Conference on Signal and Image Processing, August 2002, pp. 595-600.
2. Rossman M, Adjouadi M, Mirkovic N, Ayala M, Jayakar P, and Yaylali I, An Integrated Approach to Localize Epileptic Foci Using Relative SPECT Subtraction, Proceedings of the IASTED International Conference on Modeling and Simulation, ISBN: 0-88986-337-7, Palm Springs, CA, USA, February 24-26, 2003, pp. 342-347.
3. Rossman M, Adjouadi M, Ayala M, Yaylali I, An Interactive Interface for Seizure Focus Localization Using SPECT Image Analysis, Computers in Biology and Medicine. January 2006. Volume 36, Issue 1, pp. 70-88.
4. Adjouadi M, Ayala M, Cabrerizo M, Zong N, Lizarraga G, and Rossman, M. Classification of Leukemia Blood Samples Using Neural Networks, Annals of Biomedical Engineering. April 2010, Volume 38, Issue 4, pp. 1473-1482.

PATENTS

1. Correction of PLT Count Due to the Loss To Platelet Clumps. Method of Correcting Platelet Count Due to Platelet Clumping (with Dr. Shuliang Zhang and Jiuliu Lu), Patent has been submitted to United States Patent and Trademark Office and is currently pending.
2. Optical Detection of RBC Abnormalities. Optical Identification of a Population of RBC Cells that has Correlation with the Presence of Malarial Parasites (with Dr. John Riley and Liuliu Lu). Patent has been submitted to United States Patent and Trademark Office and is currently pending.