Using Probability Binning and Bayesian Inference to measure Euclidean Distance of Flow Cytometric data

استخدام توزيع الاحتمالات واستدلال بايزي لقياس المسافة الاقليدية لبيانات التدفق الخلـــوي

By

Rasha Mahmoud Abdel Salam Mohamed
Student ID: 110152

Dissertation submitted in partial fulfilment of the requirements for the degree of MSc Informatics (Knowledge and Data Management)

Faculty of Engineering & Information Technology

Dissertation Supervisor
Professor Sherief Abdallah

December 2014

DISSERTATION RELEASE FORM

Student Name	Student ID	Programme	Date
Rasha Mahmoud Abdelsalam Mohamed	110152	Information Technology	15th Dec 2014

Title

Using Probability Binning and Bayesian Inference to measure Euclidean Distance of Flow Cytometric data

I warrant that the content of this dissertation is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

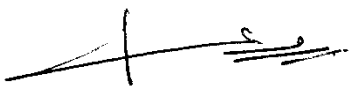
I understand that one copy of my dissertation will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make that copy available in digital format if appropriate.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my dissertation for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

Signature



Abstract

Flow Cytometry (FCM) is a microscopic technique used in many fields, especially clinical research and health care. Classical analysis of FCM data is done manually in a tedious, error prone process, which is not standardized, not open for re-evaluation and highly dependent on the experience of the analyst. Conventional analysis methods are based on comparisons of univariate or bivariate distributions for one or two channels only, while it is obvious that analyzing flow cytometric data files in a multivariate space would generate more accurate results. For this reason, many studies and researches are directed towards developing a model for automatically analyzing FCM data files, as it is difficult for human analysts to extract clear information from multidimensional data files.

The automated analysis of flow cytometric data is challenging due to many reasons especially: the unordered cells across different flow cytometric files and the features are divided across multiple FCS files for the same patient. Many approaches concentrated on resolving either the first or the second challenge, but not both of them.

In this thesis, a novel approach is introduced and validated for generating a multivariate flow cytometric data file with N-dimensions, where N is the number of the intended independent measurements. The approach was developed to resolve the main two challenges in flow cytometry – mentioned previously - using concepts of Probability Binning and Bayesian Inference.

The approach described in this thesis is validated for classifying normal and leukemia incidence cases. Also, it is validated for classifying different Leukemia types (AML, B-ALL or T-ALL). Experiments show a 100% correspondence between our results and clinical results.

الخلاصـــــة

يعد قياس التدفق الخلوي تقنية مجهرية تستخدم في العديد من المجالات، وخاصة مجالات البحث العلمي السريري والرعاية الصحية. وتعتمد بيانات التدفق الخلوي على طرق التحليل اليدوية، التي تعد طرقاً شاقة ومعرضة للخطأ، كما أنها ليست موحدة، غير مفتوحة لإعادة التقييم وتعتمد بشكل كبير على خبرة المحلل. وتعتمد طرق التحليل التقليدية على مقارنات التوزيعات وحيدة المتغير أو ثنائية المتغير لواحد أو اثنين فقط من القنوات المخصصة لقياس خصائص الجسيمات المجهرية، في حين أنه من الواضح والمنطقي أن تحليل ملفات بيانات التدفق الخلوي في فضاء متعدد المتغيرات من شأنه أن يولد نتائج أكثر دقة. وبما أنه يصعب على المحلل البشري أن يستخرج معلومات واضحة من ملفات بيانات التدفق الخلوي متعددة الأبعاد، فقد توجهت العديد من الدراسات والأبحاث العلمية نحو تطوير نموذج لتحليل بيانات التدفق الخلوي تلقائيا.

يمثل التحليل التلقائي لملفات بيانات التدفق الخلوي تحدياً لعدة أسباب، من أهمها: الخلايا والجسيمات المجهرية غير مرتبة في جميع ملفات التدفق الخلوي، كما أن خصائص وميزات الخلايا مقسمة على عدة ملفات للمريض الواحد، وليست متاحة في نفس الملف. قامت العديد من الأبحاث بوضع نهج مقترحة كثيرة لحل إما التحدي الأول أو التحدي الثاني فقط، ولكن لا يتواجد منهاجاً يقدم حلاً كاملا يشمل التحديين الأول والثاني معاً.

في هذه الأطروحة أقدم منهاجاً جديداً تم التحقق من صحته في إنتاج ملف واحد لبيانات التدفق الخلوي متعدد المتغيرات، والذي يحوي عدد (ن) من الابعاد، حيث (ن) هو عدد القنوات المستقلة والموزعة على عدة ملفات. وقد تم تطوير هذا النهج لحل التحديين الرئيسين – السابق ذكرهما- وذلك باستخدام مفاهيم توزيع خانات الاحتمالات واستدلال بايزي.

وللتحقق من صحته؛ تم تطبيق المنهاج المقدم في هذه الأطروحة لتشخيص سرطان الدم وتصنيف عدد من الحالات إلى حالات عادية سليمة وأخرى مصابة باللوكيميا. كذلك تم التحقق من صحة هذا المنهاج عن طريق استخدامه لتصنيف أنواع سرطان الدم المختلفة (AML سرطان الدم النخاعي الحاد – B-ALL سرطان الدم الليمفاوي الحاد نوع B – T-ALL سرطان الدم الليمفاوي الحاد نوع T) وقد أظهرت التجارب تطابقا بنسبة 100% بين نتائج المنهاج المطبق في هذه الأطروحة وبين النتائج السريرية .

Acknowledgements

I would like to express my thanks to all whom supported me in this thesis. First and foremost, I dedicate the deepest appreciation to my supervisor, Professor Doctor Sherief Abdalla. I would never have been able to finish my dissertation without his support, guidance and advices.

I would also like to gratefully and sincerely thank my husband and my children for their patience and support. They are always encouraging me with their love and best wishes.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Rasha Mahmoud Abdel Salam Mohamed)

Table of Contents

List of Figures

List of Tables

List of Equations

Chapter 1

1. Overview

Flow Cytometry (FC) is a widely used technique and an essential tool in many fields that are related to clinical research and health care, like Immunophenotyping, DNA analysis, microbiology and drug discovery. A flow cytometer forces thousands of cells to flow one at a time through a very thin chamber. While flowing, the physical and chemical characteristics of the cells are measured and recorded in Flow Cytometry Standard Data files (FCS). Depending on how many cell features are to be measured, more than one aliquot (for the same patient) may be analyzed, producing more than one FCS file for the same patient. (Cualing 2000), (liu et al. 2008), (Pedreira et al. 2008a), (Lakoumentas et al. 2009) and (Rowley 2014) highlight the importance of flow cytometry in Hematological Immunophenotyping in the following points:-

1. Diagnosing and monitoring blood cancer: Flow cytometers provide rapid multi-parameter antigenic fingerprinting of blast cells. Also FC is used in distinguishing lymphoid from myeloid leukemia.
2. Diagnosing and monitoring HIV/AIDS infection.
3. Detection of minimal residual disease (MRD) based on the neoplastic antigenic fingerprint, even when neoplastic cells exist at very low frequencies among majority of normal cells.

(liu et al. 2008), (Pedreira et al. 2008b) and (Bashashati & Brinkman 2009) show that traditionally, and this remains the case in many labs today, FCM data are analyzed by human experts through visualization of scatter plot arrays of different pairs of cell features, followed by gating. A gate is a selective boundary that can be used to define the characteristics of particles to include for further analysis, and eliminate unwanted particles (dead cells or debris) from results. Although gating is an important step in FCM analysis; it's considered to be a time consuming and a highly subjective process which depends mainly on the experience of the analyst.

Due to the importance of the problem and the availability of the data electronically, there has been several attempts to partially or fully automate the analysis of FCM data. However, such automation faces a number of challenges which will be explained in details in the next section.

Unordered cells over different FCS files and features distributed over different FCS files are the most important challenges. For example, in the field of diagnosing Leukemia; conventional FCS analysis methods could not be used to directly measure the distance between leukemia patients. In other words, applying traditional distance metrics to original FCS files would generate different results for the same patient and this is obviously wrong. This is due to unordered cells across different FCS files for the same patient.

1.1. Challenges facing analysis of Flow Cytometric data

(Liu et al. 2008) believe that analyzing (FCS) data files is a real challenge due to several reasons:-

1. Flow Cytometry is an example of multiparametric measurement technique that can measure up to 20 channels (attributes or features) for each cell at a time.
2. The number of cells in one (FCS) file can reach the order of $(10)^5$ or $(10)^6$. Hence the matrix of (FCS) file contains huge number of intensities of different features (channels).
3. Cells are not uniformly ordered across samples of the same patient.
4. Data of a single patient may be divided over multiple (FCS) files, depending on the technical limitations of the flow cytometer.
5. Usually in Flow Cytometry research field, the number of cases under investigation is relatively small. This produces a huge search space for the parameters of any model.

The above mentioned challenges may exist in other health-related data, but point numbers (3 and 4) represent unique challenges for flow cytometry and needs a special care.

1.2. Questions the research address

Many researches introduce different approaches to (partially or fully) automate the analysis of flow cytometric data. Generally, most approaches that attempt to represent FCM data can be divided into two main categories:-

1. Approaches addressing the challenge of unordered cells (challenge number 3 in the previous section). These approaches concentrate on analyzing separate channels (attributes) in each (FCS) data file. Vector quantization, mixture modelling and frequency binning are examples of techniques used to group similar cells in one unit and eliminate the importance of cells' order.

2. Approaches integrating channels across FCS data files. Visualization techniques, nearest neighbor classification techniques and Bayesian Inference are examples of techniques used to merge channels from different FCS files (of the same patient) into one huge FCS file.

This thesis presents a novel approach that combines two different techniques to solve the two challenges facing representation of flow cytometric data. As mentioned earlier some researches solved the first challenge, but not the second; while other researches solved the second challenge but not the first. The representation that is proposed here solves both challenges and could successfully be used on real data to achieve 100% accuracy and sensitivity in classifying Leukemia patients. The approach presented here was also used in a novel idea of 2-step classification of leukemia cases, and also achieved 100% accuracy and sensitivity in discriminating different Leukemia types, which is more challenging for human pathologists.

This thesis aims to answer the following questions:-

1. How can we represent FCM data that address the two representation challenges?
2. Can the proposed representation be used to successfully classify Leukemia cases?
3. Can the proposed representation be used to successfully classify different types of Leukemia cases?

1.3. Contributions

In this section, thesis main contributions are outlined. The following illustration shows how the thesis work answers the research questions in details.

1. How can we represent FCM data that address the two representation challenges?

(Cox et al. 1988), (Roederer et al. 2001a) and (Roederer et al. 2001b) present the concept of Probability Binning (PB) that address the challenge of unordered cells across FCS files of the same patient. (Pedreira et al. 2008c) apply Bayesian Inference (BI) to address the challenge of features divided over different FCS files, by merging histograms of different features in one global dataset. This thesis combines both Probability Binning and Bayesian Inference to present a novel approach for FCM data representation that can be used with traditional distance metrics.

2. Can the proposed representation be used to successfully classify Leukemia cases?

The proposed representation was used to generate a global (or multi-parameter) dataset that contains cases (instances) on the rows, and (binned and combined) features on the columns. This

dataset could be analyzed using different data mining techniques, and verified to accurately classify leukemia cases.

3. Can the proposed representation be used to successfully classify different types of Leukemia cases?

The representation proposed here was applied on FCS data of different leukemia types to generate a multi-parameter dataset. This dataset was then analyzed using a novel idea of 2-step classification to classify two main types of leukemia: Acute Myeloid Leukemia (AML) and Acute Lymphoid Leukemia (ALL). The representation verified to succeed in classifying different leukemia types.

Briefly, the contributions of this thesis are:

1. A representation for FCM data based on Probability Binning and Bayesian Inference.
2. First data set of leukemia cases in the UAE analyzed locally using data mining.
3. Comparative analysis of classification techniques for leukemia cases.
4. Proposing 2-stage classification of Leukemia types.

1.4. Thesis Outlines

The remaining of this thesis is organized as follows:

Chapter 2: introduces a technical description of a flow cytometer and its specifications, biological facts about different blood cells and the definition of leukemia and its types.

Chapter 3: discusses the related work to this thesis.

Chapter 4: represents the basic concepts and mathematical principals used in building the algorithm of this thesis.

Chapter 5: represents the methodology and the detailed steps of the suggested algorithm.

Chapter 6: represents the first experiment of examining (FCS) samples for the diagnosis of leukemia. The experiment includes: generating the global (FCS) dataset, using optimization and selection techniques to reduce its size, and finally analyzing it using data mining techniques. Also application of Euclidean Distance metric on flow cytometric dataset is presented here.

Chapter 7: represents the second experiment of discriminating between Acute Myeloid Leukemia (AML) and Acute Lymphoid Leukemia (ALL). The experiment includes: generating the global

(FCS) dataset, using optimization and selection techniques to reduce its size, and finally analyzing it using data mining techniques.

Chapter 8: concludes and summarizes this thesis; discusses the results and presents the future work.

Chapter 2

2. Flow Cytometry

Flow Cytometry (FC) is a widely used technique that allows simultaneous multi-parameter analysis of hundreds of thousands of individual cells. It is used to measure cell size and volume, analyze expression of cell surface and intracellular molecules and characterize different cell types in heterogeneous cell populations.

2.1. Cell preparation for Flow Cytometry

FC uses hematological and lymphoid tissues as specimens such as peripheral blood (PB), bone marrow, lymph node and thymus. Each specimen is composed of different percentages of populations of cells, characterized by cell lineage, growth stage, functioning and activation level. The identification of each cell population can be done by binding high affinity antibodies to cell's protein structure or by measuring physical cells' characteristics. In some cell populations, a simultaneous measurement of 2 or more protein structures is required to accomplish the identification process. This biological fact increases the demand for instruments with higher multi-parameter qualifications. As a result, new instruments are capable now of measuring up to 20 parameters simultaneously by developing cytometers with 20 florescence detectors as mentioned in (Bashashati & Brinkman 2009). (Pedreira et al. 2008c) and (Pedreira et al. 2013) argue that these developments in industry are still less than the requirements of clinical discrimination between several heterogeneous disease categories. For example, for accurate identification of B cell chronic lymphoproliferative disorders (BCLPDs), acute leukemia, or myelodysplastic syndromes; measuring characteristics of neoplastic cells with respect to more than 30 markers are required, while currently available cytometers does not provide 30 florescence color detectors. This potential problem of multicolor detectors limitations has been solved by staining cells with biomarkers (fluorochrome-conjugated antibodies). There are two main categories of biomarkers associated with different hematological malignancies; backbone markers and supplemental markers. The specimen is divided into aliquots and stained with different combination of markers. All aliquots are stained with backbone markers to identify the same population of cells. Different aliquots of the sample are stained with different supplemental markers for additional description of physical characteristics of populations. In other words, to solve the problem of limited number of detectors in cytometers, panels with two or more

combinations of overlapping antibodies are used to measure cells' characteristics for the complete identification of different hematological malignancies. Although staining solved the problem of florescence detectors limitations, it has encouraged the development of new automatic multidimensional analysis techniques.

2.2. Technical Description of a Flow Cytometer

(Rowley 2012) and (Pedreira et al. 2013) demonstrate the structure of Flow Cytometers (Cyto for cell, meter for measure). A flow cytometer consists of four main components: fluidic system, lasers, optics and electronic system. Figure (2) represents the main components of a flow cytometer. The test tube containing a sample is placed in the collection stage of the flow cytometer. The heterogeneous suspension in the test tube is drawn from the test tube and pumped to the flow chamber. The flow chamber allows cells to flow one at a time very quickly by Hydro Dynamic Focusing which is achieved by controlling the pressure of the sample with respect to the pressure of the sheath fluid. Figure (3) focuses on the fluidic system of a flow cytometer.

When the cell reaches the interrogation point (the point of intersection between the flow chamber and the laser light source), it is hit by a bright light laser beam. The light reflected off each cell is collected by light detectors to give information about the cell's physical characteristics. Light reflections at small angles are called Forward scatter (FS) and gives information about the size of the cell, while light reflections at large angles are called side scatter (SS) and indicates if the cell contains granules as described in (Lakoumentas et al. 2006) and (Pedreira et al. 2008b). Also cells are stained with bio markers (antibody fluorochromes) that cause certain light emission according to protein information in each cell. Light scattering and/or fluorescence emission are captured, filtered by photo-detectors and converted to electrical signals (voltage). The voltage signal is digitized and stored in a Flow Cytometry Standard (FCS) file format. The value of the voltage signal shows the intensity of light emitted from each cell. Thus each FCS data file can be described by a matrix whose columns are the channels (attributes or features of cells), and rows are the individual cells. The intersection between each row and column is the intensity of the voltage.

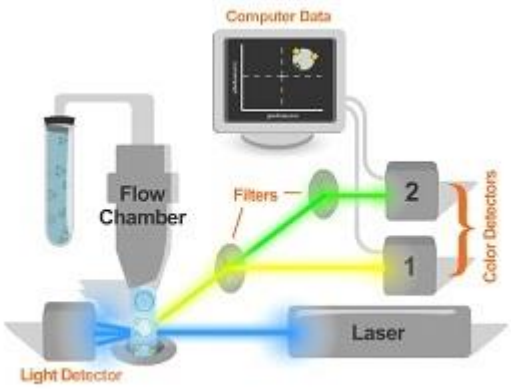


Figure 1: Flow Cytometer Structure. Adopted from (Flow Cytometry: How Does it Work? 2014)

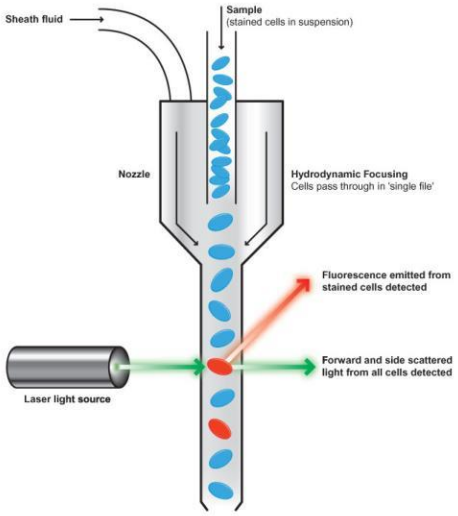


Figure 2: Cells flowing in a flow chamber of a cytometer. Adopted from (Abcam 2014)

(Pedreira et al. 2008a), (Pedreira et al. 2013) and (Lakoumentas et al. 2009) believe that many developments in FC industry have been achieved, but conventional data analysis techniques don't satisfy those complex achievements. Thus adequate developments in data representation, visualization and analysis techniques are a real challenge and an essential demand.

2.3.Different specifications of flow cytometers

There are many manufacturers for flow cytometers with different specifications. (Rowley 2012) mentions the commonly cited flow cytometers' providers in his survey. They are: BD Biosciences (its common models are: FACSCalibur, LSR II, FACSCanto and FACScan), Beckman Coulter, Dako, Guava/EMD Millipore and Miltenyi Biotec.

(Select Science 2014) shows the main differences between flow cytometers, and provides technical guidelines to choose the flow cytometer those best suites individual requirements. The following points summarize the main differences between flow cytometers:-

1- The number of parameters to be simultaneously measured is dependent on the number of lasers and detectors the flow cytometer supplied with.
2- The capabilities of the software provided with the flow cytometer. The companion software provides tools for representing the data generated by the flow cytometer in histograms, two-dimensional dot plots or some times in three-dimensional plots.
3- Gating is an important procedure in flow cytometers and varies from one manufacturer to another. Some cytometers provides manual gating, while others provide semi-automated or automated gating.
4- Some flow cytometers provide solutions to protect the samples from environmental influences.
5- The degree of automating the process of analysis varies from one manufacturer to another.
6- Flow cytometers vary according to the sensitivity, accuracy and efficiency of their fluidic, optic and electronic systems supplied with them. Figure (3) shows a detailed optic system of a flow cytometer.

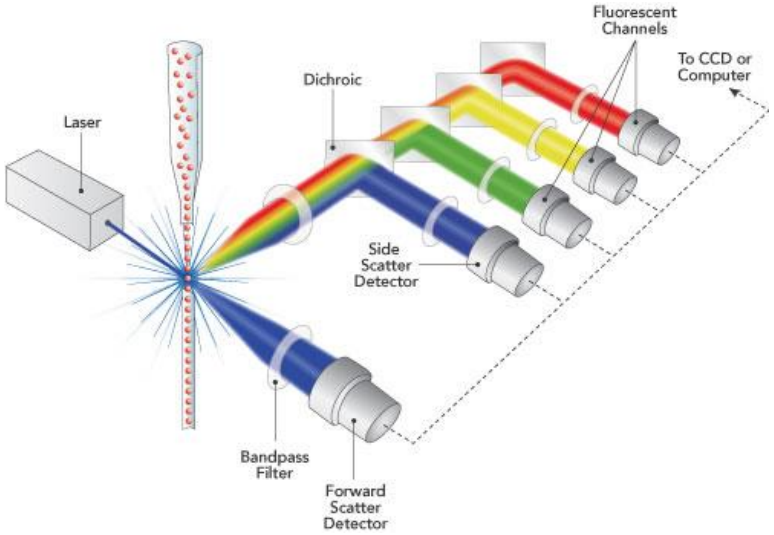


Figure 3: Optic System of FC-Adopted from (Semrock 2014)

2.4.Flow Cytometric Data files

(Pedreira et al. 2008c) and (Pedreira et al. 2013) believe that the last two decades have seen many developments in Flow cytometry. These developments included production of monoclonal antibodies, well-characterized and high quality reagents and a broad variety of compatible fluorochromes. Also great enhancements have been added to instrumentation and tools in flow cytometry, which led to:-

1. Greater number of parameters (channels) that can be simultaneously assessed for each cell.
2. Greater analysis speed of digital flow cytometers.
3. Examining tens of thousands of stained cells per seconds.
4. Generating more complex data files that contain multi-parameter distribution about millions of cells in a sample.

These developments imposed many challenges on the existing analysis tools and, thus encouraged developing new automated analysis algorithms and clear visualization tools.

Old flow cytometers used a screen attached to instrumentations to represent voltage signals emitted from each cell as analog signals in real time. Later then, computers were developed to store cells' measurements in a unique file format that distinguishes the manufacturer of the cytometer. (Pedreira et al. 2013) state that in 1984, a standardized file format was proposed to store flow cytometers data: FCS 1.0 (Flow Cytometer Standards 1.0 format). Three main formats of FCS 1.0 could be used:

1. Single FCS file format: contains information about population of cells with respect to a single parameter (single channel).
2. Double FCS file format: contains information about population of cells with respect to two parameters (two channels).
3. Multi-parameter (n dimensions) FCS file format: contains information about population of cells with respect to three parameters or more. This format is also called List Mode Files format (LMD). LMD files can be viewed as a matrix, where individual cells are on the rows of the matrix, and the parameters under investigation are on the columns of the matrix.

Since 1984, FCS 1.0 has been revised and updated to the current version of FCS 3.0.

2.5.Conventional analysis methods of FCS files

(Liu et al. 2008), (Bashashati & Brinkman 2009) and (Pedreira et al. 2008b) explain how FCS files are analyzed manually by plotting two different parameters (attributes) in a 2-D scatter plot. The resulting plot is then gated by identifying homogenous subgroups of cells of special importance to be further analyzed in another 2-D scatter plot with respect to another two channels. (Pedreira et al. 2013) demonstrate that manual technique is done by highly experienced people, who may find difficulties in extracting clear and simple information from multidimensional FCS files generated from currently available digital cytometers with many channels.

Although gating is an important step in conventional flow cytometry analysis, it is considered as a major disadvantage. (Boedigheimer & Ferbas 2008), (Liu et al. 2008), (Pyne 2009), (Bashashati & Brinkman 2009) and (Simon et al. 2012) highlight that gating is a subjective process that neglects multidimensionality of the data and miss potential subsets of cells due to projecting high dimensions' data down to a 2D space . Also the size, shape and position of the gates are highly dependent on the knowledge and experience of the analyst.

(Bashashati & Brinkman 2009) present a study performed in 15 clinical institutions. The study shows that the mean inter-laboratory coefficient of variation ranged from 17-44%. The same samples and antibodies were used, preparation of samples was standardized and analysis of data files where performed by experts in flow cytometry. Actually the study shows that high variation coefficient was accounted for gating.

2.6.Normal Blood Cells

National Cancer Institute (NCI 2013) states that all blood cells are produced by Blood Stem Cells, which exist mainly in the bone marrow. Blood stem cells can produce myeloid stem cells and lymphoid stem cells, which in turn will finally produce three types of cells:

1. White Blood Cells (WBCs) which help the body fights infection.
2. Red Blood Cells (RBCs) which carry oxygen all over the body.
3. Platelets which form clot to stop bleeding.

All blood cells have a certain life cycle which consists of many stages: formation, growth, function and death. This life cycle is controlled in the bone marrow, which will in turn produce new blood cells instead of the dead ones.

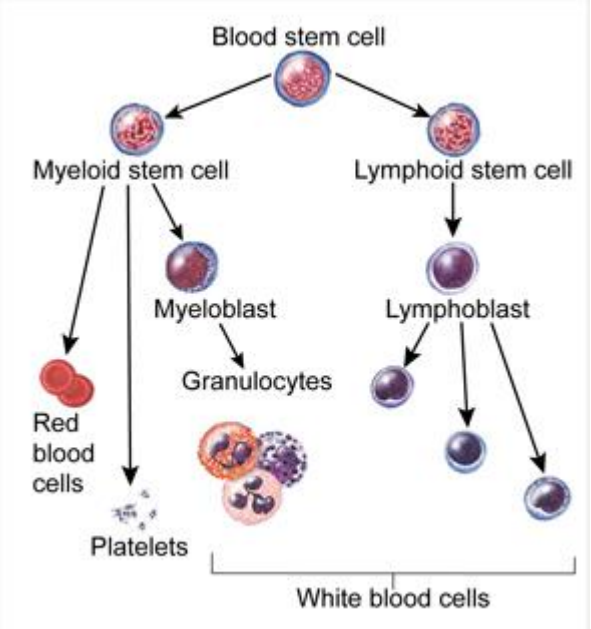


Figure 4: Generation of Normal Blood Cells - Adopted from (NCI 2013)

As shown in figure (4), WBCs can be formed from different cell lineages, Granulocytes which are generated from myeloid stem cell, and Lymphoblasts which are generated from lymphoid stem cell. Lymphoblast (Lymphocyte) cells are called Agranulocytes, and are of three types: B-cells, T-cells and Natural Killer (NK).

2.7.Leukemia, Definition and Types

(Tadmouri et al.) define Leukemia as the form of cancer that targets the blood. Another definition from National Cancer Institute (NCI 2013) Leukemia is the cancer that starts in blood-forming tissue, such as the bone marrow, and causes large numbers of abnormal blood cells to be produced and enter the bloodstream.

In Leukemia incidence cases, bone marrow produces abnormal WBCs called Leukemia and Leukemia blast cells which divide to produce copies of themselves. The copies divide again and again producing more and more leukemia and leukemia blast cells. Also in leukemia incidence cases the control of the bone marrow on the abnormal white blood cells' life cycle is disturbed resulting in undying white blood cells even if they are old or damaged. As a result, abnormal WBCs crowd out the other two types of blood cells resulting in disorders in fighting infections, controlling bleeding and delivering oxygen to tissues.

Leukemia is the most common cancer in children. It causes disturbances of functions in their immune system causing fevers and infections. Leukemia also causes disruption in generating

different blood cells, causing anemia and bleeding problems. It may also results in tumors formation due to accumulation of leukemia cells in different organs causing headache or pain.

As mentioned in the previous section, WBCs can be produced from two main cell lineages; Lymphoid and Myeloid. Also the degree of abnormality of WBCS can be strong and sudden (acute) or weak and slow (Chronic).This results in four main types of Leukemia- as explained in (NCI 2013)- they are:-

1. Acute myeloid leukemia (AML) affects myeloid cells and grows quickly. Leukemic blast cells collect in the bone marrow and blood. AML is most common in adults.
2. Acute lymphoblastic leukemia (ALL) affects lymphoid cells and grows quickly. Leukemic blast cells usually collect in the bone marrow and blood. (Tadmouri et al.) state that ALL is the most common leukemia in children.
3. Chronic myeloid leukemia (CML) affects myeloid cells and usually grows slowly at first. Blood tests show an increase in the number of white blood cells. The abnormal white blood cells work almost as well as the normal white blood cells. There may be a small number of leukemic blast cells in the bone marrow. CML is most common in adults.
4. Chronic lymphocytic leukemia (CLL) affects lymphoid cells and usually grows slowly. Blood tests show an increase in the number of white blood cells. The abnormal cells work almost as well as the normal white blood cells. CLL is most an adult disorder.

Chapter 3

3. Related Works

(Pedreira et al. 2008c) presents a novel algorithm for analyzing flow cytometric data files for the diagnosis and identification of B-cell chronic lymphoproliferative disorders. The algorithm is based on generating a single flow cytometric data file from many flow cytometric data files that contain information of events (cells) stained with different combinations of antibodies. The final file should present information about all antibodies for each individual cell. The algorithm uses a 3D vector of the common parameters (FSC, SSC and CD19) to represent all cellular events in different aliquots. Then a nearest neighbor approach is used to estimate the intensity value of different parameters for each individual cell. In other words, the approach tries to guess the intensity value of all parameters for all cells in a sample depending on a common 3D frame and nearest neighbor approach. For example, suppose two aliquots A and B with parameters (FSC, SSC, CD19, Xa) for aliquot A and parameters (FSC, SSC, CD19, Xb) for aliquot B. The common parameters are used to build the 3D vector that works as a general visualization structure. Now, the main goal is to estimate the intensity value of Xa for all cells in aliquot B, and estimate the intensity value of Xb for all cells in aliquot A, in such a way to generate a super file that contains information about (FSC, SSC,CD19,Xa,Xb). This goal is achieved by using the nearest neighbor approach which-for a cell in B- finds the nearest neighbor cell in A, and assigns its Xa value to that cell in B. Note that Xa was not directly measured for cells in B. The algorithm show high agreement between the results obtained from analyzing individual flow cytometric data files and the results obtained from analyzing the super global flow cytometric file that contains (infinite) number of parameters.

(Simon et al. 2005) develop a clustering algorithm based on cell intensities in all dimensions (channels) at once. The algorithm depends on applying successive clustering techniques to flow cytometric data files. First, K-means clustering is applied to all original data; in order to distinguish between interesting data cells, and those particles without biological importance. Then another hierarchical clustering phase is applied on a random subset of observations. The result of this phase is just a start to build a more accurate Gaussian model. Although this model shows successful results; but it depends on an expert knowledge to exclude unimportant clusters in the early stages of the algorithm.

(Liu et al. 2008) apply important preprocessing steps by converting flow cytometric data files from its original structure showing intensity value for each cell at each channel to another structure showing cell counts for every intensity value at each channel. This intensity distribution structure makes it easy to directly compare flow cytometric data files. To overcome the problem of huge number of features; regression analysis is used. As the number of features is significantly reduced; data files could be separately clustered at each channel. The evaluation of algorithm shows that a great enhancement has been done to the quality and efficiency of reduced features clustering compared with original features clustering.

(Costa et al. 2006) and (Pedreira et al. 2008b) present successful approaches to classify different neoplastic B-cell disorders depending on the concept of Vector Quantization (VQ). VQ is a technique built on modelling of probability density functions by a number of prototype vectors. Each vector is representing an –approximately- equal number of data points closest to it. Each vector is viewed as the centroid of group of data points surrounding it, as in K-means clustering. This approach reduces the number of features in the search space, as only one vector is representing a group of data points. This clustering technique was used at each channel separately and could successfully classify different neoplastic B-cell disorders depending on classifying lymphocytes.

(Lakoumentas et al. 2006), (Boedigheimer & Ferbas 2008), (Chan et al. 2008) and (Lakoumentas et al. 2009) represent similar approaches for automatically gating lymphocytes. The approach depends on the concept of mixture model. In this model, each data point is assumed to be generated from a mixture of probability distributions. For each data point, the model can guess the most probable distribution to which it belongs. This model is sometimes referred to as Bayesian Clustering, as it uses the concept of Bayesian networks in clustering data points.

(Shih et al. 2013) develop a graphical based approach, in which healthy cases, B-Chronic Lymphocytic Leukemia (B-CLL) cases and Follicular Lymphoma (FL) cases are plotted in a 3-D 5-parameter model. The 5 parameters of this model are the five biomarkers: CD5, CD10, CD19, Kappa and Lambda. The model is based on clinical facts that normal cases show positive values for CD19, but negative values for both CD5 and CD10, with approximately equal populations expressing Kappa and Lambda. On the other hand CLL cases show positive values for CD19 and CD5, but negative values for CD10. Lastly, FL cases show positive values for CD19 and CD10, but negative values for CD5. In both CLL and FL populations express either Kappa or Lambda

light chains. The model is built by training on FCM data cases, and then used to fit the algorithm on test cases. The graphical nature of the approach shows great efficiency and easiness in diagnosing different B-Lymphocytosis disorders.

(Pedreira et al. 2008a) represent a probabilistic approach on BCLPD (B-cell chronic lymphoproliferative disorders). The approach applies Bayes Theorem to predict the probability of every data point (event) either to belong to a normal population or a neoplastic population. Assuming that events follow Gaussian distribution functions; conditional probabilities of an event drawn from normal or neoplastic case can be easily calculated. Thus this paper represents a direct application of Bayesian Law.

(Aghaeepour et al. 2013) represent FLOWCAP. Flow Critical Assessment of Population is a software developed to compare the performance of different flow cytometric data analysis techniques. The software compares the performance with respect to manual expert results, and with respect to actual clinical outcomes. The paper aims to develop techniques for automatic analysis of flow cytometric data files, and to provide guidelines about accurately using those techniques.

(Bashashati & Brinkman 2009) represent a general framework for analyzing FCS data files. Majority of papers in this filed focus on automatic gating techniques (selecting a homogenous subset of cells which share one function or have similar interesting characteristics). Other papers focus on applying supervised or unsupervised learning techniques in finding correlations between cells' characteristics and clinical results. Supervised learning algorithms are used when there is a label variable (attribute) that can distinguish between the events in a data set. While in unsupervised learning all the variables are the same, and the goal here is to separate the events of the data set into similar groups or clusters where data points (events) of one cluster are most similar to each other than events in other clusters. In most cases, unsupervised learning techniques are more suitable to analyze flow cytometry data files, as usually we don't have a prior knowledge about the nature of cells and other particles in blood samples.

(Zare et al. 2010) argue that spectral clustering is the best technique for analyzing FCS data files. However, direct application of spectral clustering on FCM data sets (of size 300,000 events for example) will consume extraordinary time and memory (approximately two years! and 5 terabytes of memory). Thus reducing data events first would solve the problem. The approach is based on the following steps:-

1. Plotting data points (vertices) on a grid
2. Running faithful sampling to produce much less data points where spectral clustering can be performed easily.
3. The lost information due to sampling is compensated by adding weighted edges to data points produced in step (2). The edge is a connection between two data points (vertices) with a weight (thickness) appropriate to the density of the region, i.e. a denser region is weighted with thicker edge. This way, the lost information about the density of data points will be retrieved.
4. Spectral clustering is applied on the modified data set.

The overall previously mention algorithm is called SamSPECTRAL, which was tested and verified to be the first successful application of spectral clustering on large datasets.

Chapter 4

4. Basic concepts

This research presents a novel approach to represent FCS data. This representation could be used with traditional distance metrics. The procedure presented here is built upon two main concepts:

1. Probability Binning.
2. Bayesian Inference (to generate Joint Probability distribution).

(Pedreira et al. 2008c) present a novel algorithm to generate a multivariate distribution from different FCS files. The algorithm - explained previously- is similar to the algorithm presented here in their main objective, which is merging different parameters measured for different aliquots into one flow cytometric file. The main difference between the former algorithm and the one presented here; is the nature of FCS files being used for merging. (Pedreira et al. 2008c) use the raw FCS files which presents intensity values for individual cells for each channel (parameter) measured, while the procedure presented here uses intensity histograms (intensity distribution values) for each channel measured. The currently presented procedure is much simpler as it neglects the order of cells, and cares only for their quantity at each discrete intensity value.

4.1. Cox Method and Theory of Probability Binning

(Cox et al. 1988) state that comparing two or more multivariate dataset histograms can determine if the datasets differ from one another in terms of their respective fluorescence intensity distributions, however, the current statistical tests do not identify the exact region of difference on the fluorescence intensity scale. For example, the Kolmogorov-Smirnov test computes a P value to assess the similarity of two cell populations in terms of their fluorescence intensities but the test does not provide where on the intensity scale the difference lies. Identifying these "difference regions" requires modifying the comparison test to show gates of difference and accommodate limited intensity ranges which can drop down to 5% of the sample population. This limitation of current tests promoted the use of Cox Method.

(Cox et al. 1988) introduce Cox Model for identifying the regions of difference -on intensity scale- between samples. In order to understand how Cox Method is applied, suppose that we have two stained blood samples (Control sample) and (Test sample). Cox Method group cells in

equally sized bins according to their florescence intensities. Each bin has unequal number of cells. These bins are then compared between the two samples to detect the regions with significant difference in cell counts on the intensity scale. Although the size of bins in Cox Method is arbitrary, it affects the statistical power of the comparison. If the bin is too narrow, it will have small cell counts and lead to less powerful statistics. If the bin size is too wide, it will have large cell counts and the resolution of intensities is –therefore- lost.

(Roederer et al. 2001a) and (Roederer et al. 2001b) introduce the theory of Probability Binning (PB) to provide a metric that determines the degree of similarity or dissimilarity between two or more multivariate distributions. It can be used to discriminate samples and classify them according to their biological differences. Probability binning is just a modified version of the original algorithm named Cox Method (or Cox Model).

Probability Binning theory (PB) uses unequal sized bins with equal number of cells in each bin. This will result in sufficient amount of cells in each bin, in such a way to provide powerful statistics of difference between control sample and test sample. This method suffers from losing resolution at unpopulated areas of florescence intensity scale; as it may group outliers with inliers together in one bin to complete the pre-specified amount of cells in each bin. Thus we can say that equal frequency binning is biased against outliers, which may represent neoplastic cells.

(Roederer et al. 2001b) describe and validate the Probability Binning algorithm to compare univariate distributions according to their florescence intensity. A control univariate distribution is divided into bins with equal frequencies in each bin. This algorithm is considered a mini-max approach as it minimizes the maximum variance of the control dataset. The same bins are then applied to a test distribution. As in Kolmogorov-Smirnov (K-S) statistics, a chi-squared P values are calculated to assess the similarity of control and test distribution on the basis of bin – by – bin comparison. PB algorithm uses a Monte-Carlo simulation for chi-squared P values and converted them to another metric T(x) score. The metric T(x) scales with the degree of similarity or dissimilarity between two distributions, which can't be decided with P values of K-S test for example. This metric can also be used to determine the relative distance between different distributions and a control dataset.

Following the same concept, (Roederer et al. 2001a) show that PB algorithm can also be applied for comparing multivariate distributions, where each bin is made up of hyper-rectangles of n dimensions, where n is the number of parameters (attributes) to be simultaneously measured.

Again, each bin must contain the same number of events (cells), and the same number of bins is applied on both control and test data sets. As mentioned previously, PB algorithm could generate a metric T(x) that can rank distributions according to their similarity or dissimilarity to a control distribution. The algorithm was applied to multivariate immunophenotyping data, and verified its ability to discriminate different distributions and rank them according to their biological meaningful differences.

4.2.Probability, conditional probability and Bayes Law

A probability is a numeric value representing the possibility that a certain event will occur. (Duda et al. 1999) express the concept of probability as a random variable X that can assume a finite number m of different values v_i in the space of X. Thus:

$$P(x) = v_i, \text{ for } i=1, 2, 3\ldots m$$

Equation 1: Probability of a random variable (X)

Where *x* is the value that the random variable (X) can take and m is the number of different values that (X) can take.

An impossible event is the event that will never occur, i.e. has a possibility value=0.

$$\mathrm{P(x)} = 0 \quad \text{For impossible event}$$

Equation 2: Probability of an impossible event

On the other hand a certain event is the event that of sure will occur, i.e. has a possibility value=1.

$$\mathrm{P(x)} = 1 \quad \text{For certain event}$$

Equation 3: Probability of a certain event

(Duda et al. 1999) suggest expressing the set of all possible probabilities that a variable will take in terms of a probability mass function $P(x)$. The probability mass function must satisfy these two conditions:-

$$\mathrm{P(x)} \geq 0$$

Equation 4: Condition 1 on Probability mass function

$$\sum_{\mathrm{x}} \mathrm{P(x)} = 1$$

Equation 5: Condition 2 on Probability mass function

Conditional Probability is the probability that event X will occur given that -or on the condition that- event Y has already occurred, and is denoted by P(X|Y). We can find P (X|Y) by consider the sample space truncated to just the event Y, and is given by:

$$P(X|Y) = \frac{P(X\ and\ Y)}{P(Y)}$$

Equation 6: Conditional Probability (X|Y)

Similarly, the conditional probability P(Y|X) can be found by considering the sample space truncated to just the event X, thus:-

$$P(Y|X) = \frac{P(Y\ and\ X)}{P(X)}$$

Equation 7: Conditional Probability (Y|X)

Note that P (X|Y) ≠ P (Y|X).

As mentioned in (Andrade 2010) and using (**Error! Reference source not found.**) and (**Error! Reference source not found.**), Bayes Law is derived:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Equation 8: Bayes Law

Equations (Equation 6: Conditional Probability (X|Y)), (Equation 7: Conditional Probability (Y|X)) and (Equation 8: Bayes Law) will be used in Chapter 5 (Methodology), to derive the main joint probability equation that merges between the different channels from different FCS files.

Chapter 5

5. Methodology

The diagnosis of various hematological diseases, such as different types of lymphocytic leukemia and lymphoma, requires comprehensive data analysis of lymphocytes. (Lakoumentas et al. 2006) confirm that it is necessary to discriminate lymphocytes from all other leukocytes for the diagnosis of leukemia. The study agrees that using normal FS/SS (Forward Scatter/Side Scatter) gating is not enough in the discrimination process. Depending on the Forward Scatter FS attribute to discriminate cells according to their relative size is not always successful because cells are usually mixed. Also depending on the SS attribute to discriminate cells according to their granularity (complexity) is not always successful as SS can distinguish granulocytes, but can't distinguish lymphocytes and monocytes.

(Lacombe et al. 1997) introduce the fact that leukemic blast cells express low values of the leukocyte common antigen (CD45) marker, while normal lymphocytes and monocytes express high values of CD45. They compare between normal FS/SS gating (without using any immunological marker) and CD45/SS gating. The results of their study show that the former gating produces overlapping areas between blast cells and normal cell populations, while gating on CD45/SS can precisely discriminate between blast cells and normal cells. (Lacombe et al. 1997) show that using CD45 plus two or three lineage-specific markers can achieve good discrimination between leukemic blast cells and normal cells. Their procedure was applied on AML cases, but also succeeded for different types of acute leukemia, even those which have low percentages of blast cells.

Moreover, (Lakoumentas et al. 2006) point out that CD3 and CD19 are two antigen markers that can discriminate between different subtypes of lymphocytes: B-cells and T-cells, as they can produce distinctive flow cytometric attributes.

(Lacombe et al. 1997) and (Lakoumentas et al. 2006) assure that CD19 antigen marker can recognize B-cell lymphocytes, while CD3 antigen marker can recognize T-cell lymphocytes.

However, the required antigen attributes, CD3 and CD19, are not always available in the same flow-cytometric file of the same patient examination. Therefore, this thesis suggests adding the common antigen attribute CD45 to help in merging the channels, as CD45 is available in almost

all flow cytometric files of the same patient and is also used for discriminating between lymphocytes among other leukocytes.

Therefore, and depending on the above mentioned facts, this thesis claims that using CD45 with CD3 and CD19 give excellent phenotypic determination of the leukemic blast cells, and hence provides accurate leukemia diagnosis.

The procedure presented here aims at merging intensity histograms of three channels into single multidimensional histogram. The single multidimensional histogram will contain intensities of all possible combinations of the binned channels. This integration process is done for each patient, and then all multi-dimensional histograms are gathered in one global dataset. The columns of the global dataset introduce all combined- binned channels, and the rows represent the cases.

This chapter introduces the three main stages of the procedure. The first section presents the modified algorithm of probability binning, the second section presents Bayesian Inference that will be used to merge the channels and the third section describes how the multi-parameter dataset will be generated to achieve the proposed representation of FCS data.

Sample (FCS) data files are provided in LMD file format (listmode) from Pathology Section at a hospital in Dubai. Exploration of the data and statistical calculations were performed in the statistical package R [(1)] version 3.0.2.

5.1.Stage (A): Applying the modified Probability Binning algorithm

Stage (A) is the step where normal (FCS) files (of cell's intensity distributions) are converted to (intensity histograms). This stage is similar to Cox method and Probability Binning algorithm, except in the size of bins and the amount of cells in each bin. The algorithm presented here is much simpler and easier than the former two algorithms, and also verified its ability to generate high accuracy results. The following steps describe this stage: -

1. Define the possible range of intensities (minimum and maximum intensity values) by exploring the backbone channels (CD3, CD19 and CD45) in (FCS) files.
2. The range of intensities is divided into bins of unequal size with no condition on the amount of cells in each bin. Sizes of bins are selected by trial and error, taking into account striking the balance between compact representation and accuracy of results.

[(1)] R is a free software environment for statistical computing and graphics. R is available for free download at http://www.r-project.org/. R version 3.0.2 released on 25-9-2013

Generally, exploring (FCS) files shows that intensity histograms of channels under consideration are J-shaped with left peak, so small bin sizes were selected at the peak and near it, while larger bin sizes are chosen beyond the peak. The same binning is applied on the three channels.

3. Intensities of cells in individual channels are categorized and grouped in the corresponding bins, i.e. intensity of each cell is categorized according to the selected bins' sequence, and placed in the appropriate bin.

Figures (5, 6 and 7) show intensity distributions of cells stained with backbone markers CD3, CD19 and CD45 respectively. Each distribution contains (60690) cells. Note that these plots are sensitive to cells flow order. While figures (8, 9 and 10) show histogram plots of intensities for the same cells. These plots are independent of cells' flow order, and generally following a J-shaped curve with left peak. Also note that all histogram plots have the same binning sequence.

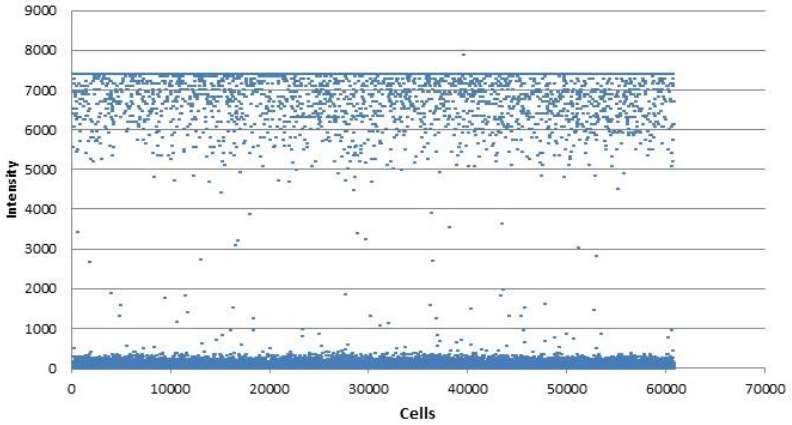


Figure 5: Cells' Intensity Distribution at channel CD3

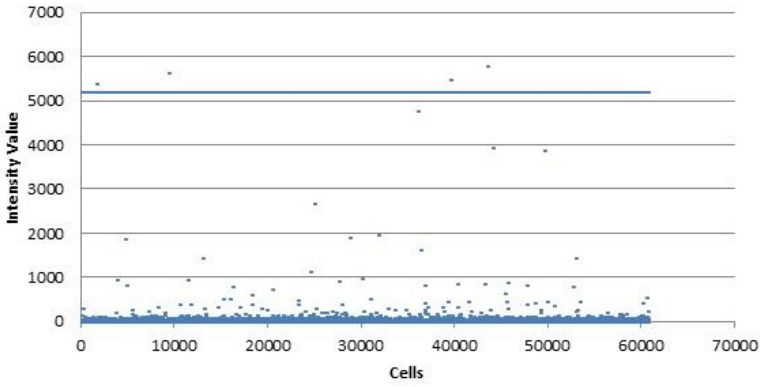


Figure 6: Cells' Intensity Distribution at channel CD19

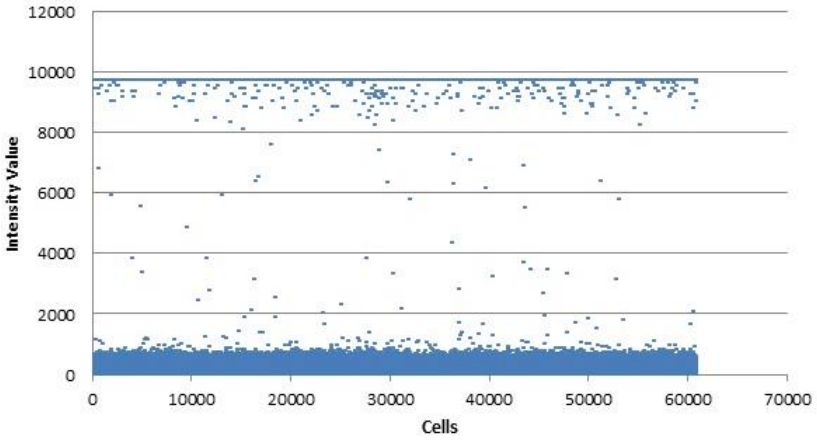


Figure 7: Cells' Intensity Distribution at channel CD45

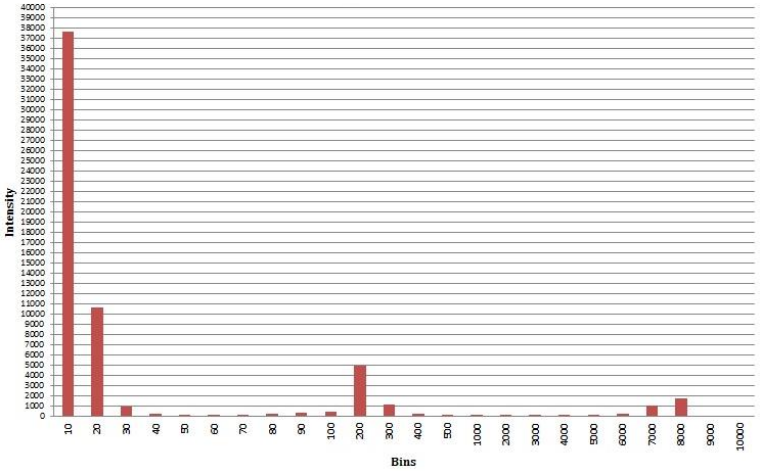


Figure 8: Intensity Histogram for channel CD3

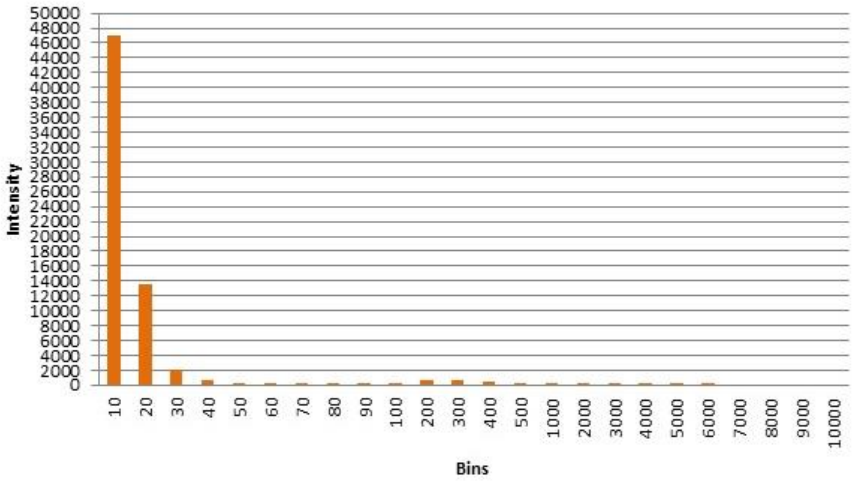


Figure 9: Intensity Histogram for channel CD19

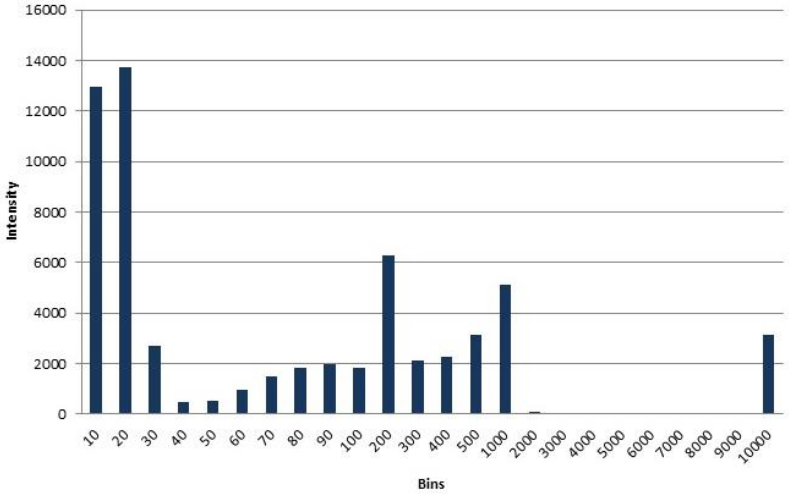


Figure 10: Intensity Histogram for channel CD45

5.2. Stage (B): Using Bayesian Inference to generate the joint probability equation of the channels

Bayesian Inference is used to join histograms across different FCS files. Suppose we have two FCS files representing two aliquots of the same patient. The FCM analysis of the first aliquot generated the FCS file with the following features: (B_1, B_2…..B_n, S_{11}, S_{12}…..S_{1m}). The FCM analysis of the second aliquot generated the FCS file with the following features: (B_1, B2…..B_n, S_{21}, S_{22}…..S_{2m}). Note that the features (B_1, B2…..$B_{n)}$ are the backbone markers that are common across all aliquots, while features (S_{i1}, S_{i2}…..$S_{im)}$ are supplemental markers that are specific to the i^{th} aliquot. Let H1 and H2 be the histograms of the first and second FCS file respectively. Now to calculate the joint histogram of all features (B_1, B_2…..B_n, S_{11}, S_{12}…..$S_{1m,}$ S_{21}, S_{22}…..$S_{2m)}$, we

assume that probabilities associated with supplemental features are independent given the backbone features. Recall joint probability equation is given by:

$$P(X,Y) = P(X|Y)P(Y)$$

Equation 9: Joint Probability Equation

Therefore:

$$H= P(B_1, B_2\ldots..B_n, S_{11}, S_{12}\ldots..S_{1m,}\, S_{21}, S_{22}\ldots..S_{2m}) = P(S_{11}, S_{12}\ldots..S_{1m} \mid B_1, B_2\ldots..B_n)\, P(S_{21}, S_{22}\ldots..S_{2m})\, P(B_1, B_2\ldots..B_n)$$

Equation 10: The joint histogram of all features

Substituting using (Equation 6: Conditional Probability (X|Y)), we get:

$$H = \frac{P(S11, S12 \ldots.. S1m, B1, B2 \ldots.. Bn)\; P(S21, S22 \ldots.. S2m,\; B1, B2 \ldots.. Bn)}{P(B1, B2 \ldots.. Bn)}$$

Equation 11: The joint histogram of all features

Consider the two supplemental channels CD3 and CD19 are independent (as they are obtained from different FCS files of the same patient examinations), and CD45 is the backbone channel. Substitute in (Equation 11) to get:

$$P(CD3, CD19, CD45) = \frac{P(CD3, CD45)P(CD19, CD45)}{P(CD45)}$$

Equation 12: Joint Probability Equation of channels CD3, Cd19, CD45

From equation (2) we can say that the joint histogram (the joint probability of all features) can be calculated from the bivariate probability distribution between (CD3, CD45), bivariate probability distribution between (CD19, CD45) and the univariate probability distribution of (CD45).

5.3. Stage (C): Generating contingency tables for (CD3, CD45), (CD19, CD45) and frequency table for (CD45)

In this stage, bivariate distributions and univariate distributions are generated in order to perform equation (2). The following steps describe this stage:-

1. From stage (A): cells' intensity distributions are converted into intensity histograms for CD3, CD19 and CD45 separately.

2. Combine frequency histograms of both CD3 and CD45 in a contingency table. Do the same for CD19 and CD45. These contingency tables describe the relationship between (CD3, CD45) and (CD19, CD45) as bivariate data distributions.
3. Calculate frequency table for CD45 as a univariate data.
4. Substitute in (Equation 12: Joint Probability Equation of channels CD3, Cd19, CD45)
5. Stages (A and C) are repeated for all cases. The resulted combined attributes of all cases are then used to generate a multi-parameter dataset.

Note that CD3 and CD19 are usually not in the same FCS file,

Note that CD3, CD19 or CD45 can be repeated in different examinations (FCS files) for a single patient, so steps 1, 2 and 3 can be performed for those channels (attributes) from any (FCS) file that contains them, as all (FCS) files for the same patient will show the same final results. Appendix (A) shows the detailed steps of the algorithm presented here on one sample case.

A clear limitation of the approach presented here is the explosion of the generated combined attributes. The number of the generated combined attributes (N) is exponential in the number of individual attributes. N is given by:

$$N = (b)^{m}$$

Equation 13: Number of generated combined attributes

Where:-

m: is the number of individual features (both supplemental and backbone channels).

b: is the number of break points used in binning.

This limitation can be resolved by using effective manual binning, accurate selection of individual features and applying feature reduction techniques on the generated (Combined Attributes) dataset.

Chapter 6

6. Discriminating between normal and abnormal leukemia cases

In this chapter, the first and second questions of this thesis will be answered:

1. How can we represent FCM data that address the two representation challenges?
2. Can the proposed representation be used to successfully classify Leukemia cases?

This chapter presents experiments to discriminate between normal and abnormal leukemia cases. The chapter consists of two main parts: generate the proposed representation of FCS data by generating the combined attributes dataset (CA), then analyzing the resulted (CA) dataset using Rapid miner 5.3[(2)] to classify leukemia cases.

6.1. Generating the Combined Attributes (CA) dataset

Generating the combined attributes dataset requires calculating the joint probability distribution of the three channels (CD3, CD19, CD45) by calculating equation (2) for all the available cases (30 normal cases and 5 Leukemia incidence cases). Recall equation (2)

$$P(CD3, CD19, CD45) = \frac{P(CD3, CD45)P(CD19, CD45)}{P(CD45)} \qquad (1)$$

Where:-

P(CD3, CD19, CD45): is the joint probability of the combined backbone attributes (CD3, CD19, CD45).

P(CD3, CD45): is the joint probability of two channels (CD3, CD45).

P(CD19, CD45): is the joint probability of two channels (CD19,CD45).

P(CD45): is the discrete probability of one channel (CD45).

The procedure described previously was evaluated using data of 30 normal cases and 5 abnormal cases. The data files of the cases were provided in LMD (Listmode) file format. FCS Express 4

[(2)] Rapid Miner 5.3: is a data mining program. It is available as a stand-alone application for data analysis and as a data mining engine. Rapid Miner is available for download at: http://sourceforge.net/projects/rapidminer/files/latest/download

Plus- Research Edition [3] was used to convert files into FCS file format and CSV file format. .The three channels are discretized using 24 break points at: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, which results in 25 manually-set bins [4]. Therefore, the maximum number of combined attributes that will be generated in the dataset is $(24)^3$= 13824 attributes. The actual generated dataset contains 6175 attributes, as the remaining attributes have zero value for all cases, so they were cancelled out to simplify the analysis and reduce the size of the dataset.

Case	label	10 10 10	10 10 100	10 10 1000	10 10 20	10 10 200	10 10 2000	10
10n	normal	0.047716	0.01082	0.031198	0.017657	0.149289	0.01866277	0.0
11n	normal	0.182279	0.003812	0.110022	0.032442	0.077145	0.033266556	0.0
13n	normal	0.020818	0.028015	0.046507	0.008782	0.327855	0.017901391	0.
14n	normal	0.430916	0.00194	0.129535	0.029279	0.014221	0.060895363	0.0
15n	normal	0.395545	0.00113	0.092632	0.039585	0.033393	0.066066468	0.0
16n	normal	0.638405	0.003592	0.013913	0.062682	0.039192	0.01250938	0.0
17n	normal	0.154798	0.018963	0.0045	0.100683	0.059297	0	0.0
1a	abnormal	0.075154	0.000285	0.085094	0.001942	0.003776	0.041237664	0.0
1n	normal	0.209148	0.018059	0.006772	0.185833	0.057737	0	0.0
20n	normal	0.024587	0.007153	0.014074	0.003278	0.14557	0.005944004	0.0
21n	normal	0.252474	0.008481	0.028805	0.027178	0.110421	0.042952	0.0
22n	normal	0.467225	0.000571	0.018773	0.001756	0.013764	0.00224931	0.0
23n	normal	0.004292	0.028211	0.004088	0.003374	0.141395	7.32E-05	0.0
24n	normal	0.038547	0.018346	0.006125	0.021437	0.185962	0.004830918	0.0
25n	normal	0.608796	0.001389	0.078749	0.019321	0.009781	0.05985739	0.0
26n	normal	0.274721	0.004254	0.051588	0.124176	0.052375	0.039747628	0.0

Figure 11: Part of the combined attributes (CA) Dataset

Figure (11) shows a part of the generated (CA) Dataset, were the cases are on the rows; the attributes including a label attribute are on the columns. The values of numerical attributes show the joint probability of the combined channels (CD3, CD19 and CD45). For example, the combined attribute [10 10 10] means the joint probability to have CD3∈ [0, 10], CD19∈ [0, 10] and CD45∈ [0, 10]. This attribute equals 0.0477 for case 10n, while it equals 0.4672 for the case 22n. Also, note that attribute [10 10 2000] equals 0 for case 17n. This means that the probability to find CD3 ∈ [0, 10], CD19 ∈ [0, 10] and CD45 ∈ [1000, 2000] equals 0 for that case.

[3] FCS Express 4 Plus is a flow cytometry software package designed and produced by De NOVO software. A demo version is available at http://www.denovosoftware.com/site/demo-overview.shtml.

[4] More sophisticated binning methods can be used, but this intuitive manual setting worked just fine.

Note that abnormal cases has the ID of the form (number + a), while normal cases has ID of the form (number + n).

6.2. Analyzing the (CA) dataset using Rapidminer 5.3

Rapid Miner5.3 is used to build seven different classifier models for the (CA) Dataset. All models are built with default Rapid Miner5.3 parameters unless otherwise stated. Cross validation (with 10 folds and stratified sampling) is used to measure the performance of the classifier model. The Cross Validation operator in Rapid Miner5.3 is a nested operator that has two sub-processes; training and testing. In 10 folds cross validation, the (CA) dataset is divided into 10 subsets of equal size. In the training sub-process, 9 subsets are used to build the model and the last 10^{th} subset is used in the testing sub-process to measure the performance of the model on unseen data. This process is repeated 10 times, each time with different subsets for training and testing. The final performance of the model is the average of its performances in all 10 cycles. Table (1) shows the performance measures for each classifier model.

#	Model	False Positive	False Negative	Accuracy (%)	Precision (%)	Recall (%)
1	Decision Trees	1	1	95	80	80
2	Rule Induction	1	1	95	80	80
3	Logistic Regression	7	3	72	22	40
4	K Nearest Neighbor	2	2	89	60	60
5	Perceptron	2	1	93	67	80
6	Naïve Bayes	2	1	92	67	80
7	Support Vector Machines	7	3	72	22	40

Table 1: Classification Models for (CA) dataset

It is clear that Decision Tree (DT) and Rule Induction (RI) classifiers gained the highest accuracy of 95%, followed by Perceptron with 93%, then Naive Bayes (NB) with 92%, followed by K Nearest Neighbor (KNN) with 89%, and finally the lowest accuracy of 72% was acquired by Logistic Regression (LR) and Support Vector Machines (SVM).

Decision Tree and Rule Induction models use the combined attribute [10 1000 3000] to classify normal and abnormal cases. Figures (12 and13) show a scatter plot diagram for all cases versus the combined attribute [10 1000 3000] in both DT and RI models. it is clear that the combined attribute [10 1000 3000] in both DT and RI could not classify all cases correctly due to the relatively small number of cases with respect to the huge number of combined attributes. Also note that two abnormal cases and one normal case - instances located inside the rectangular area in figures 14 and 15- have relatively close values of the combined attribute [10 1000 3000], hence both classifiers could not correctly classify all cases based on that attribute only.

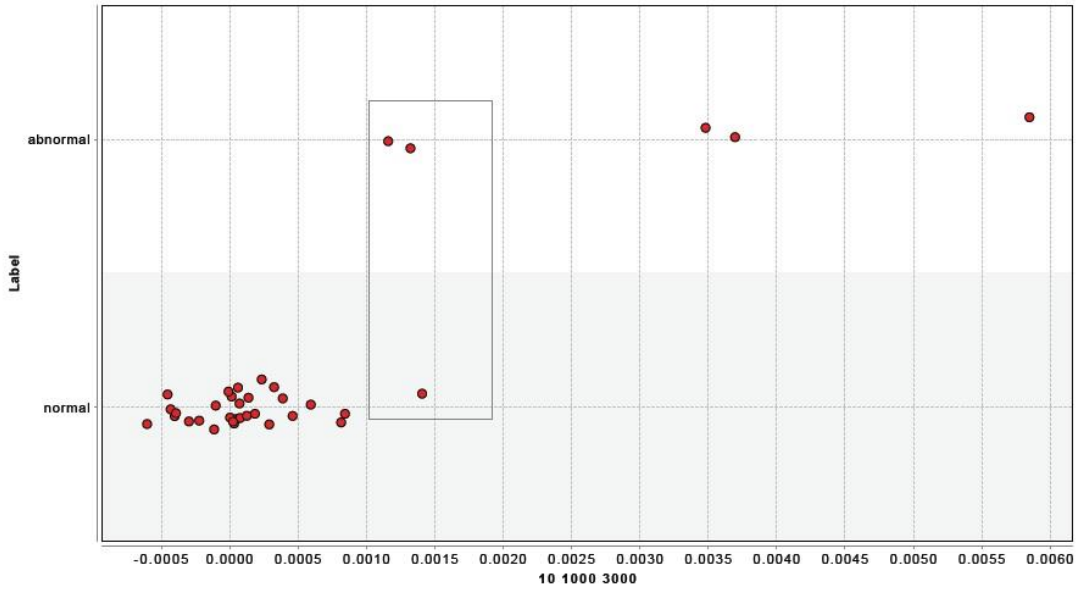


Figure 12: The Combined Attribute [10 1000 3000] Vs Label in Decision Tree Classifier

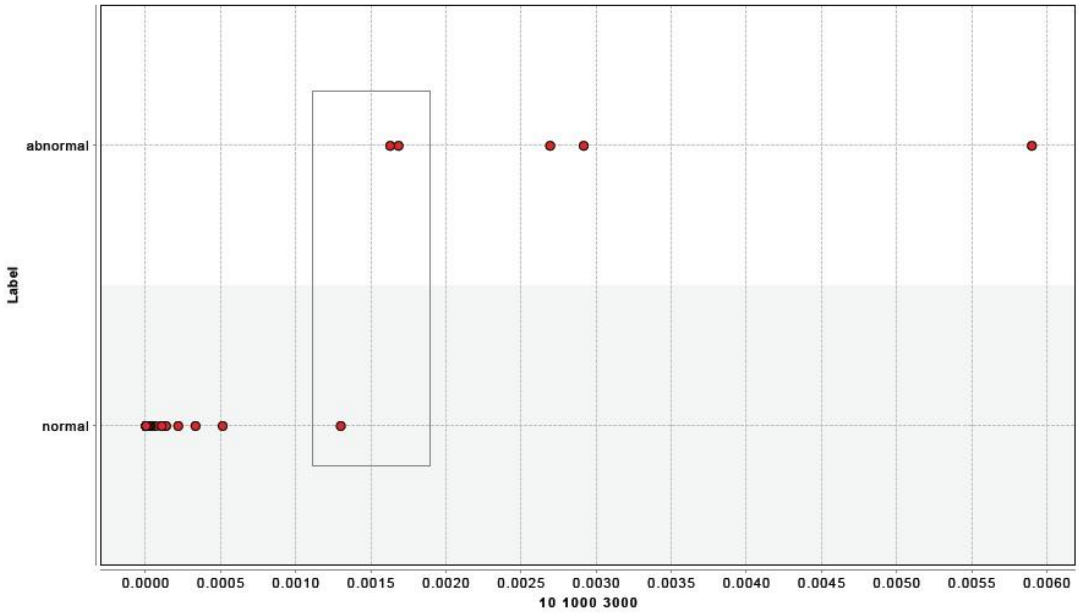


Figure 13: The Combined Attribute [10 1000 3000] Vs Label in Rule Induction Classifier

In Logistic Regression model, the combined attribute [10 400 2000] achieved the highest weight of (0.032344), followed by the combined attribute [10 500 2000] with a weight of (0.031455). The third highest weight of (0.027805) was assigned to the combined attribute [10 300 2000].

In Perceptron model, the highest three weights of (0.02295), (0.007493) and (0.005515) were assigned to the three combined attributes [10 200 1000], [10 300 2000] and [10 300 1000] respectively.

In Support Vector Machine model, the highest weight of (0.002302) was assigned to the combined attribute [10 500 2000], followed by the attribute [10 400 2000] with a weight of (0.002263), and the third highest weight of (0.001977) was assigned to the combined attribute [10 1000 3000].

6.3.False Positive and False Negative Errors

This medical experiment is considered as a binary classification problem, where the result of classification is either normal or abnormal. In this case, errors can occur when the classifier model indicates that an instance is normal while it is not, or indicates that an instance is abnormal while it is not. These two errors are called false positive and false negative, respectively. A false positive error exists when a test result shows that a condition is present but it is not in fact presented, while a false negative error exists when a test result shows that a condition is not present, but it is in fact presented. Although both errors affect the results of the medical testing

problem, but the medical case investigated here is a testing problem of life-and-death situation. In false positive, the patient may need to do extra tests and medical diagnosis to ensure the validity of the results. This will cost some time, money and effort. While in case of false negative, the patient may not receive the required treatment based on his actual case, so it will cost his/her life! In this particular problem, it is better to add unnecessary tests instead of endangering someone's life! Thus it is essential for the algorithm to be practical; to reduce false negative errors to zero.

In the following sections, different solutions will be introduced that may increase the accuracy of classification and reduce the number of false negative errors to zero.

6.4. Apply feature selection techniques

The resulted (CA) dataset consists of 6175 attributes and only 35 cases (events). This is an example of many domains that have tens or hundreds of thousands of features and only few training examples. (Guyon & Elisseeff 2003) state that gene selection from micro array data and text categorization are two famous examples of such cases. The gene selection case is very similar to our task here, where a typical classification task requires separating healthy patients from cancer patients depending on their gene expression profile. The dataset usually has less than 100 patients as training examples, while individual profile consists of around 6000 – 60,000 variables. The variables are coefficients corresponding to abundance of mRNA in a sample, which may be hematological lymphoid tissue biopsies.

As mentioned previously, the huge number of attributes compared to the number of training samples, may cause “Curse of Dimensionality “ which may lead to low classification accuracy. This ensures that there are many redundant and irrelevant variables in the resulted dataset. This is also the case in gene selection classification task. Actually there are many benefits behind suggesting the approach for feature selection, they can be summarized in the following points:-

1. Enabling data visualization and simple representation.
2. Reducing the required training and testing time.
3. Handling the problem of “Curse of Dimensionality”.
4. Improving prediction accuracy.

Many feature selection techniques use variable weighting scheme as a basic step. This scheme gives higher weights to variables that could correctly classify the training samples.

(Aggarwal 2014) classifies feature selection techniques to: filter models, wrapper models and embedded models. In Filter method, a ranking criterion on one feature or a subset of features is used to evaluate their classification ability. The filtering of variables is done as a preprocessing step, and the result is irrespective of the used machine learning technique. In wrapper method, a model uses an iterative classification algorithm with built in-feature selection technique. In each iteration; the classification algorithm evaluates a subset of features for their predictive ability. If the accuracy of prediction is improved, then the embedded subset of features is selected, otherwise, it is neglected. Due to embedding feature selection model in the classification algorithm, the resulted subset of features is dependent on the used classification algorithm. For example, SVM classification would choose features that linearly separate the two classes, while nearest neighbor classifier would choose features that cluster classes in spherical regions.

As examples on applying feature selection techniques; section 6.5 presents attribute weighting process, and section 6.7 presents application of a wrapper model.

6.5. Using Attribute Weighting

An attribute weighting process is established in Rapid Miner 5.3. The process consists of three operators: (1) Retrieve CA dataset Operator, (2) Weight by SVM operator, then (3) Select by Weight operator. The operator (weight by SVM) is used to calculate the relevance of each attribute from the input dataset with respect to the class attribute. In this case the attribute weights are the coefficients of a hyper-plane calculated by a SVM classification model. The third operator is used to select only those attributes that satisfy a pre-specified condition. The condition specified here is "select Top 3 attributes". The process resulted in the following attributes:-

1. Attribute [10 500 2000] with a weight of 1.
2. Attribute [10 400 2000] with a weight of 0.983.
3. Attribute [10 1000 3000] with a weight of 0.859.

The first attribute [10 500 2000] has a weight of 1, which means that a SVM hyper-plane has a classification accuracy of 100% using that attribute, while, using the second attribute will result in a classification accuracy of 98%. Constructing a hyper-plane at the last attribute has the least classification accuracy which is 86%.

If the same process repeated again, but with the operator (Weight by Chi-square statistic), chi-squared statistics will be calculated to measure the relevance of each attribute from the input dataset with respect to the class attribute. The results are shown below:

1. Attribute [10 400 2000] with a weight of 1.
2. Attribute [10 500 2000] with a weight of 1.
3. Attribute [10 1000 3000] with a weight of 0.844.

The first and second attributes have a chi-square statistic (p) value of 1, which means 100% significance with respect to the class attribute, or in other words, the first and second attributes can be used to classify all the data points with accuracy of 100%.

6.6. Constructing a (3D) visualization Model

The top three combined attributes acquired from the previous section can be used to visualize all instances in a 3D space. Figures (14 and 15) show a 3D model of all cases, where X-axis is the combined attribute [10 400 2000], Y-axis is the combined attribute [10 500 2000] and Z-axis is the combined attribute [10 1000 3000]. Figure (16) shows a 2D model of all cases in a 2D space, where X-axis is the combined attribute [10 400 2000] and Y-axis is the combined attribute [10 500 2000].

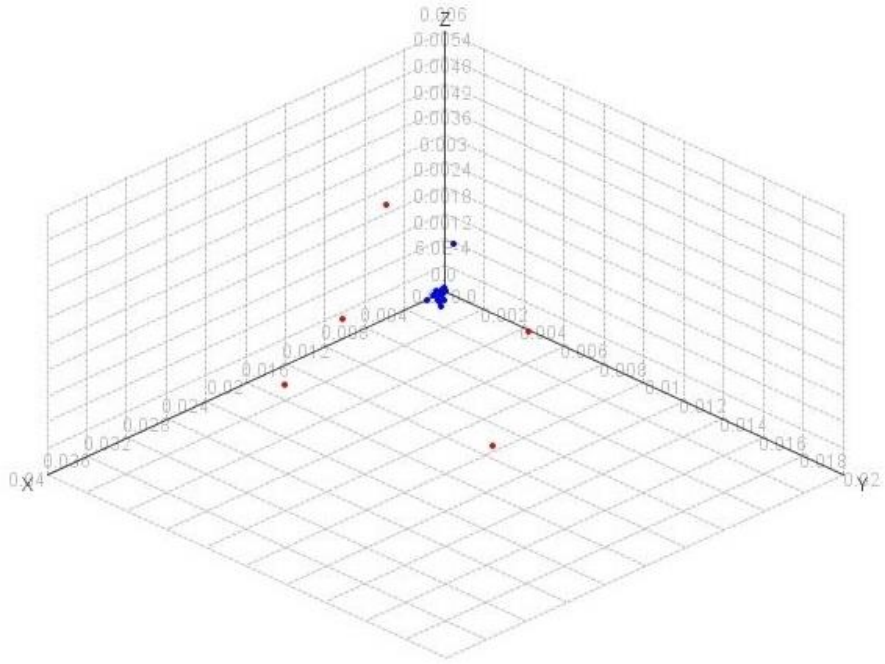


Figure 14: A 3D Visualization Model (X-Y-Z)

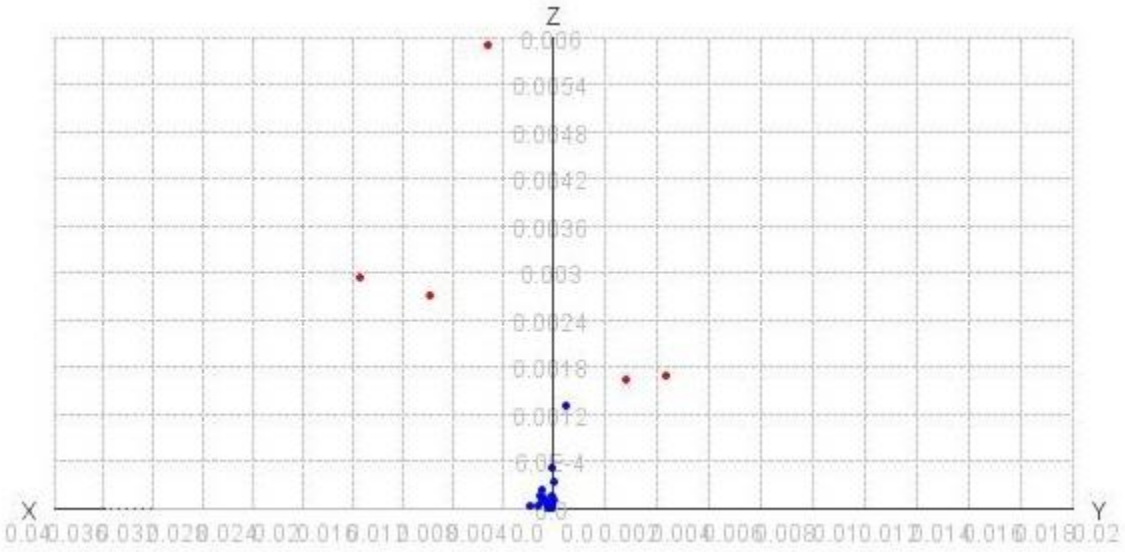


Figure 15: A 3D Visualization Model (X-Y-Z)

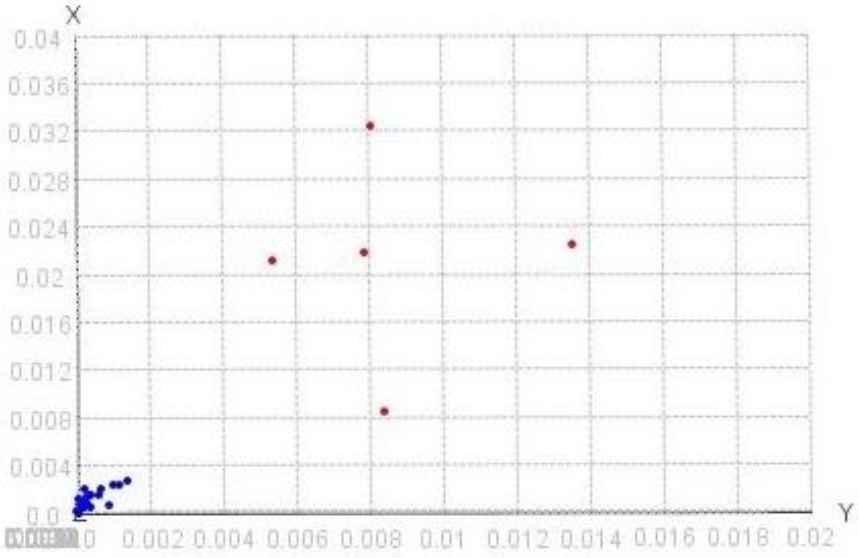


Figure 16: A 2D Visualization Model (X-Y)

6.7.Apply Wrapper Model

Rapid Miner 5.3 is used to establish a wrapper model. The model consists of a main process and two sub processes. The main process is used to weight the combined attributes from CA dataset (weights by SVM or Chi-Square Statistics), and optimize the selection of attributes according to

their weights. The first sub-process is used to build a DT classifier model using the optimized attributes. The second sub-process is used to measure the performance of the classifier using a 10 folds cross validation with stratified sampling.

Table (2) shows the results of using a wrapper model with SVM as a weighting operator, while table (3) shows the results of using a wrapper model with Chi-Square Statistics as a weighting operator.

Using SVM, a hyper-plane is constructed at the attribute [10 500 2000] and it ranks the cases according to their similarity to a control case by weights. If the weight is high, it means that a case is dissimilar to a control case, while a small weight value means the case is similar to a control case. In table (2) the highest weights are given to abnormal cases, which are the most dissimilar to a control case. All normal cases are assigned small weights, which mean they are very similar to the control case. The cases 17n, 1n and 8n have a rank of (0) which means they coincide with the control case.

As shown in table (3), similar results have been acquired when using chi-squared statistics, which utilized the attribute [10 400 2000] to rank the cases. Again cases 17n and 1n coincide with the control case. The case 5a is the most dissimilar to the control case.

#	Test Case	Label	10 500 2000
1	5a	abnormal	0.013582169
2	4a	abnormal	0.008391017
3	3a	abnormal	0.008100031
4	2a	abnormal	0.007879461
5	1a	abnormal	0.005397414
6	27n	normal	0.001383272
7	4n	normal	0.001124756

8	11n	normal	9.73041E-4
9	31n	normal	8.44496E-4
10	25n	normal	6.39992E-4
11	24n	normal	5.75109E-4
12	13n	normal	3.87695E-4
13	6n	normal	3.506E-4
14	28n	normal	3.11119E-4
15	15n	normal	2.76114E-4
16	9n	normal	2.61165E-4
17	16n	normal	2.38633E-4
18	29n	normal	2.29637E-4
19	22n	normal	1.98977E-4
20	7n	normal	1.98017E-4
21	14n	normal	1.89732E-4
22	5n	normal	1.61201E-4
23	33n	normal	1.02683E-4
24	20n	normal	4.97E-5
25	21n	normal	4.75E-5
26	10n	normal	4.02E-5
27	30n	normal	3.86E-5

28	2n	normal	3.63E-5
29	3n	normal	1.11E-5
30	23n	normal	9.15E-6
31	26n	normal	9.05E-6
32	32n	normal	4.89E-6
33	17n	normal	0.0
34	1n	normal	0.0
35	8n	normal	0.0

Table 2: A wrapper model with SVM weighting operator

Figure (17) shows a scatter plot describing the results of the wrapper model with SVM operator; while figure (18) shows a scatter plot describing the results of the wrapper model with Chi-Squared statistics operator. Note the color indicator in both figures ranks from dark blue to red. This indicator ranks the cases according to their value of the selected combined attribute: (attribute [10 500 2000] in (SVM) and attribute [10 400 2000] in (Chi-Squared)). Normal cases (dark blue) have low values and are the most similar to a control case, while abnormal cases have higher values. The red instance in both figures has the highest rank; thus considered the most dissimilar to the control case.

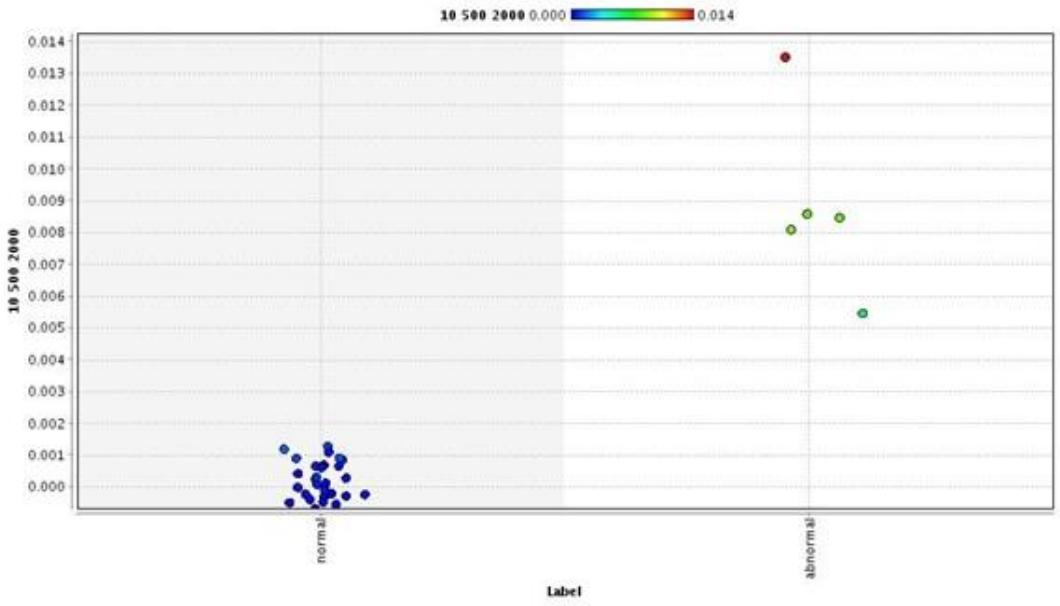


Figure 17: A scatter Plot of a Wrapper Model with SVM weighting operator

ID	Case	Label	10 400 2000
1	3a	abnormal	0.032198994
2	5a	abnormal	0.022323338
3	2a	abnormal	0.02150929
4	1a	abnormal	0.02101963
5	4a	abnormal	0.008285248
6	27n	normal	0.002621629
7	11n	normal	0.002266613
8	4n	normal	0.002262591
9	14n	normal	0.002043932
10	25n	normal	0.001913939

11	24n	normal	0.001509662
12	28n	normal	0.001485752
13	6n	normal	0.001460835
14	15n	normal	0.001271796
15	7n	normal	0.001153161
16	30n	normal	0.001095223
17	29n	normal	7.50147E-4
18	9n	normal	6.87276E-4
19	5n	normal	6.44805E-4
20	31n	normal	6.16421E-4
21	16n	normal	5.43204E-4
22	13n	normal	5.05689E-4
23	22n	normal	3.37397E-4
24	26n	normal	2.98718E-4
25	33n	normal	2.22479E-4
26	2n	normal	9.6911E-5
27	20n	normal	6.62E-5
28	3n	normal	3.88E-5
29	10n	normal	2.68E-5
30	8n	normal	2.47E-5

31	23n	normal	2.14E-5
32	21n	normal	1.19E-5
33	32n	normal	8.55E-6
34	17n	normal	0.0
35	1n	normal	0.0

Table 3: A wrapper model with Chi-Squared weighting operator

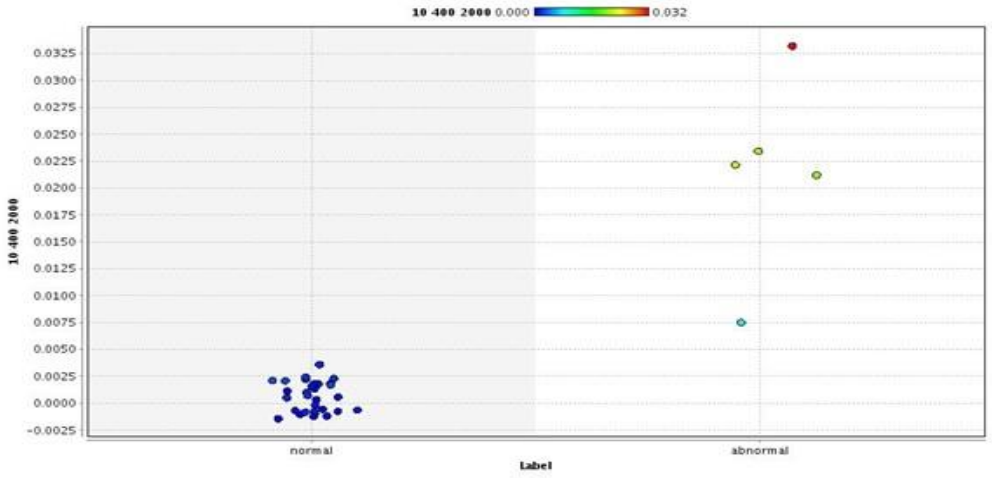


Figure 18: A scatter Plot of a Wrapper Model with Chi-Squared weighting operator

6.8. Analyzing the Reduced Dataset using Rapid Miner 5.3

Using the results gained from the previously applied feature selection and optimization techniques, we can construct a reduced dataset which consists of three combined attributes and 35 instances. The three selected combined attributes are: X is the attribute [10 400 2000], Y is the attribute [10 500 2000] and Z is the attribute [10 1000 3000].

Table (4) shows the results of analyzing the reduced (CA) dataset using Rapid Miner 5.3. Comparing table (1 and 4), it is clear that the performance of all models has enhanced significantly due to using the reduced dataset instead of the (CA) dataset. Also two models (Naïve Bayes and Support vector Machines) achieved 100% accuracy; which means that those

models have a hyper-plane that could excellently separate between normal and abnormal instances.

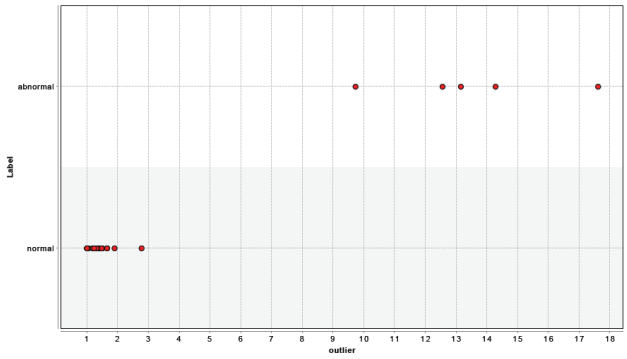
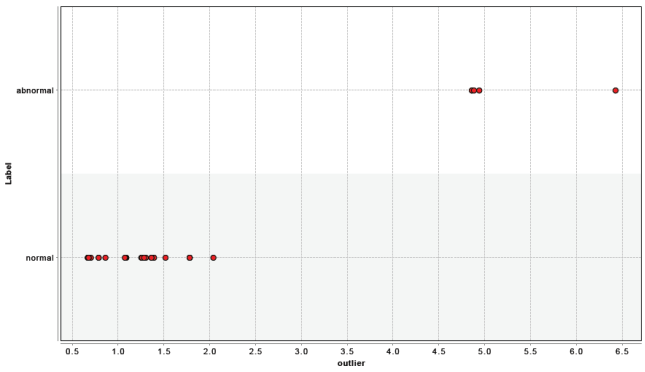
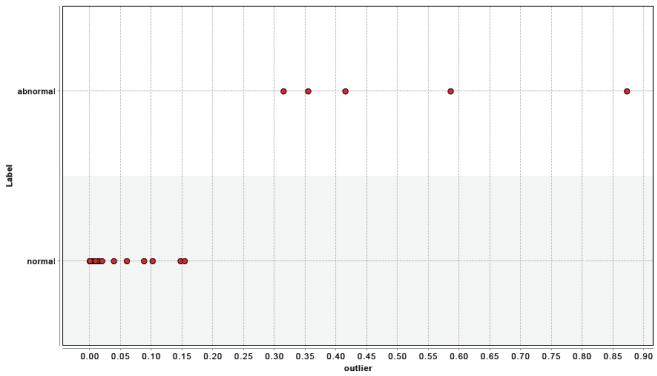
#	Model	False Positive	False Negative	Accuracy (%)	Precision (%)	Recall (%)
1	Decision Trees	1	0	97.5	83.3	100
2	Rule Induction	1	0	97.5	83.3	100
3	Logistic Regression	0	1	97.5	100	80
4	K Nearest Neighbor	0	1	97.5	100	80
5	Naïve Bayes	0	0	100	100	100
6	Support Vector Machines	0	0	100	100	100

Table 4: Classification Models for the reduced (CA) dataset

6.9. Anomaly detection techniques

Anomaly detection is the problem of finding instances in data that don't conform to a normal expected behavior. Those non-conforming instances are called outliers or anomalies. Anomaly detection problem has a special importance in many domains like fraud detection in banking credit cards, intrusion in networking, military surveillance for enemy activities and diseases in medical research. (Chandola et al. 2009) mention many supervised techniques that are used in detecting anomalies in the field of medical research and diagnosis. Neural Networks, Bayesian Networks, Rule-Based systems, parametric statistical modeling and Nearest Neighbors techniques are examples of supervised techniques used in anomaly detection especially in medical and health domain.

Table (5) shows the results of applying nearest neighbors based techniques and statistical based techniques in detecting outliers (abnormal instances) in the reduced dataset, which has three combined attributes X, Y and Z, and 35 instances; of which 30 are normal and 5 are abnormal.

#	Anomaly Detection Technique	Outlier Score (Normal)	Outlier Score (Abnormal)	Plot View (Outlier Score Vs Label)
1	Nearest Neighbor Based:(LOF) (K=10)	<3	>9	
2	Nearest Neighbor Based:(COF) (K=20)	<2.5	>4.5	
3	Nearest Neighbor Based: LoOP (K=20)	<0.2	>0.3	

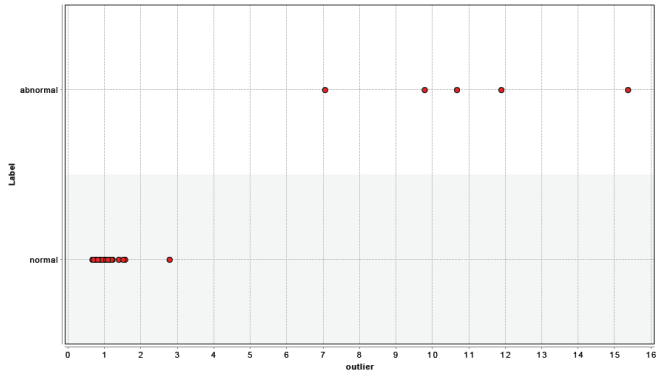
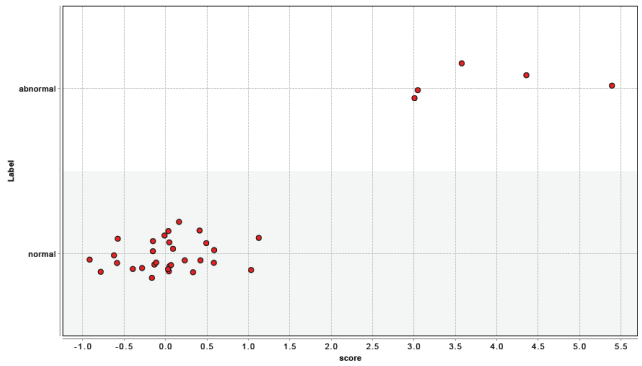
4	Nearest Neighbor Based: INFLO (K=10)	<3	>7	
5	Statistical Based: (HBOS)	<1	>3	

Table 5: Anomaly Detection Techniques

Note that (LOF) is Local Outlier Factor technique, (COF) is Connectivity Outlier Factor technique, (LoOP) is Local Outlier Probability technique and (INFLO) is Influenced Outleirness technique. These techniques are nearest neighbor based and are sensitive to K (the number of instances to be considered in the neighborhood of each individual instance). (Chandola et al. 2009) conclude that all Nearest Neighbors (NN) based techniques are built on one assumption that inlier instances occur in dense neighborhoods, while outliers occur far from their closest neighbors. Thus these techniques need to define a distance or similarity measure between two pairs of instances. The choice of the similarity measure is dependent on the type of attributes. For example, Euclidean distance is used with continuous attributes, while simple matching coefficient is used with categorical attributes. The last technique mentioned in table (5) HBOS is Histogram Outlier Score which is a statistical based technique.

6.10. Measuring Euclidean distance

(Duda et al. 2012) assert that Euclidean distance expresses the degree of similarity (or dissimilarity) between samples in a data set. Euclidean distance is the most commonly used

distance metric and is defined as the geometric distance in a multidimensional space. Euclidean distance in 3 dimensions' space can be calculated by the following equation:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$$

Equation 14: Euclidean Distance between two points p and q in a 3D space

Where d is the Euclidean distance between two instances p and q.

(Roederer et al. 2001a) and (Roederer et al. 2001b) mention that p value of chi-squared test measures the similarity between the control case and any test case. If p has a low value, this means the test case is very similar to the control case, and vice versa, if p has a high value, this means the test case is dissimilar to the control case. Using this fact and recalling tables (2 and 3), we can consider case (1n) as the control case, as it has a p value of (0). The three dimensions of the space are: X, Y and Z. Table (6) shows the Euclidean distance between case (1n) - as a control case - and all other cases.

#	**Test Case ID**	**Euclidean Distance**	**Label of Test case**
1	3a	0.033333	Abnormal
2	5a	0.026181	Abnormal
3	2a	0.023654	Abnormal
4	1a	0.021868	Abnormal
5	4a	0.011911	Abnormal
6	27n	0.002982	Normal
7	4n	0.002578	Normal
8	11n	0.002467	Normal
9	14n	0.002053	Normal
10	25n	0.002021	Normal
11	31n	0.001665	Normal
12	24n	0.001615	Normal
13	28n	0.001524	Normal

14	6n	0.001518	Normal
15	15n	0.001303	Normal
16	7n	0.001175	Normal
17	30n	0.001096	Normal
18	29n	0.000785	Normal
19	9n	0.000735	Normal
20	5n	0.000665	Normal
21	13n	0.000645	Normal
22	16n	0.000593	Normal
23	22n	0.000413	Normal
24	26n	0.0003	Normal
25	33n	0.000245	Normal
26	2n	0.000103	Normal
27	20n	8.92E-05	Normal
28	21n	5.04E-05	Normal
29	10n	5.01E-05	Normal
30	3n	4.04E-05	Normal
31	8n	2.47E-05	Normal
32	23n	2.33E-05	Normal
33	32n	9.92E-06	Normal
34	17n	0	Normal

Table 6: Euclidean Distance Between The control Case (1n) and all test cases

Table (6) is arranged in a descending order. Notice that abnormal cases are the most dissimilar with respect to the control case (1n) because they have the highest Euclidean distance ranging

from (0.033) to (0.011). On the other hand, normal cases have lower Euclidean distances with respect to the control case, ranging from (0.0029) to (0). Also note that case (17n) has a (0) Euclidean distance, which means that it coincides with the control case, and this is what we have revealed previously in tables (2 and 3).

Chapter 7

7. Discriminating between Acute Myeloid Leukemia (AML) and Acute Lymphoid Leukemia (ALL)

In this chapter, the third question will be answered: Can the proposed representation be used to successfully classify different types of Leukemia cases?

This chapter presents experiments to discriminate between acute myeloid leukemia (AML) and acute lymphoid leukemia (ALL) using both concepts of Bayesian Inference and Binning. As mentioned earlier; binning is used to change intensity values into categories (Bins), and Bayesian Inference is used to generate the joint probability distribution that merges backbone channels from different FCS files into one multi-parameter FCS file.

The above question will be answered by:

1- Representing FCS data of the selected channels and generate the combined attributes (CA) dataset.
2- Analyzing CA dataset using Rapidminer 5.3.

As mentioned earlier, lymphocyte cells can mainly be divided into B-cells and T-cells, thus (ALL) cases can be subdivide into B-ALL and T-ALL. This chapter also includes experiments on further classifying (ALL) cases into either B-ALL or T-ALL.

7.1. Generating the Combined Attributes (CA) dataset

(Van Dongen et al. 2012) present backbone attributes used for diagnosing different hematological malignancies in cooperation with INFINICYT [(5)] software. Using the backbone channels of AML, B-ALL and T-ALL, we can generate a multi-parameter dataset that contains combined attributes for diagnosing different leukemia incidence cases.

The following table shows the backbone channels for AML, B-ALL and T-ALL as mentioned in (Van Dongen et al. 2012):

[(5)] Infinicyt is software for analysis and interpretation of flow cytometry acquisition files. Infinicyt is available at: http://www.infinicyt.com/

Type of Leukemia	Backbone markers		
AML	CD45	CD34	CD117
B-ALL	CD45	CD34	CD19
T-ALL	CD45	cyCD3	CD3

Table 7: Backbone channels for diagnosing AML, B-ALL and T-ALL

The approach was evaluated using data of (4) AML cases, (4) B-ALL cases and (3) T-ALL cases. The data files of the cases were provided in LMD (Listmode) file format. They were converted using INFINICYT software into FCS and CSV file formats. Exploration of the data and statistical calculations were also performed in the statistical package R version 3.0.2 released on 25-9-2013.

7.2.Exploring the cases and generating CA dataset

The same procedure[(6)] of binning and generating the joint probability distribution between channels can be applied here using the 6 backbone channels (CD45, CD34, CD117, CD19, CD3 and cyCD3), but it will definitely be more complex due to the increased number of channels. It is clear that number of generated attributes is exponential in the number of cell features. Thus, if we could reduce the number of features, the resulting combined attributes (CA) dataset will be of reasonable size. The question now is: how can we select the correct set of attributes?

By browsing FCS files of all cases, it was found that CD45 is the backbone channel which is available in all FCS files. Also it was found that CD117, CD19 and cyCD3 is not available in some cases. Therefore, the supplemental channels that are available for all cases are only CD34 and CD3. Thus the procedure is applied on the set of channels: CD34, CD3 and CD45. The following steps summarize the procedure presented in this thesis:-

1. The backbone channels are discretized by 11 break points, they are: -100000, -10000, 0, 1000, 5000, 10000, 50000, 100000, 300000, 500000 and 1100000, which result in 12 bins.
2. Applying the joint probability equation (2) on the new selected backbone channels:

[(6)] The procedure is described in details in Chapter 5 (Methodology) and in (Appendix A).

$$P(CD34, CD3, CD45) = \frac{P(CD34, CD45)P(CD3, CD45)}{P(CD45)}$$

Equation 15: Joint Probability Equation of channels CD3, CD34, CD45

The maximum expected number of combined attributes is $(11)^3$=1331. The actual combined attributes dataset has only 617 attributes, because the remaining attributes of zero value for all cases were deleted to reduce the (CA) dataset size and simplify the analysis. Figure (19) shows a part of the generated global (CA) dataset.

Cases	Label	0 0 0	0 0 1000	0 0 10000	0 0 100000	0 0 300000	0 0 5000	0 0 50000	0 0 500000	0 1000 0	0 1000
1AML	AML	4.38E-05	0.000164	2.97E-05	0.000933733	0.001649899	1.70E-05	0.001405195	3.58E-05	0.000256	0.001
1B	ALL	7.86E-06	0.000202	8.22E-05	0.002884245	5.94E-06	7.72E-05	2.44E-05	1.11E-05	0.000539	0.000
1T	ALL	6.16E-05	0.000561	8.09E-06	4.22E-05	0.000936371	5.22E-06	3.28E-05	0	0.000118	0.001
2AML	AML	0	5.04E-05	2.93E-05	0.000626985	0.000464864	4.63E-06	0.001083862	0	0	0.00
2B	ALL	7.55E-05	0.001712	0.000211364	0.000524891	5.33E-05	0.001034496	0.000306952	0	3.49E-05	0.001
2T	ALL	4.86E-06	6.47E-05	0	0.003403583	0.001060647	5.12E-06	0.000947626	0	4.86E-05	0.000
3AML	AML	2.12E-05	0.000149	2.90E-06	0.000499791	0.003135926	1.12E-05	0.00028038	5.29E-06	0.000123	0.001
3B	ALL	8.90E-06	0.000154	2.76E-05	0.000236753	1.68E-05	5.42E-06	0.00026284	0	2.92E-05	0.000
4AML	AML	0	4.31E-05	5.85E-06	0.000214314	0.000536105	4.01E-06	6.24E-05	1.40E-05	2.22E-05	0.000
4B	ALL	0.000196	0.001281	6.35E-05	0.00022754	2.95E-05	7.68E-06	4.87E-05	0	0.000187	0.00
4T	ALL	0.000113	0.000822	2.80E-06	0.00033286	0.00040484	1.20E-05	0.000210181	3.49E-06	0.001026	0.010

Figure 19: A part of the generated (CA) dataset.

The first column (Cases) is cases ID, the second column (label) is the label of the case (either ALL or AML), and the rest of columns are the combined attributes. For example, the column (0 0 5000) is the joint probability of the combined attributes that has CD34 ∈ [0, 1000], CD3∈ [0, 1000] and CD45∈ [5000, 10000].

7.3.Analyzing the (CA) dataset using Rapid Miner 5.3

Rapid Miner 5.3 was used to analyze the (CA) dataset using different classifiers. The table below shows the results of these classifiers. It is clear that (DT), (KNN) and (Perceptron) achieved the highest accuracy of 70%, which is still a low accuracy for the sensitive problem in our hands. A possible reason for this low accuracy is the large number of combined attributes with respect to the small number of instances (cases).

#	Model	False AML	False ALL	Accuracy (%)	Precision (%)	Recall (%)
1	Decision Trees	2	2	70	71.43	71.43
2	Rule Induction	2	4	50	55.56	71.43
3	Logistic Regression	3	3	45	57.14	57.14
4	K Nearest Neighbor	3	1	70	80	57.14
5	Perceptron	0	3	70	70	100
6	Naïve Bayes	1	3	65	66.67	85.71
7	Support Vector Machines	2	3	55	62.5	71.43

Table 8: Classification Models for CA dataset

7.4.Attribute Weighting techniques

An attribute weighting process is established in Rapid Miner 5.3. The process consists of three operators: (1) Retrieve CA dataset Operator, (2) Weight by SVM operator, (3) Select by Weight operator. The operator (weight by SVM) is used to calculate the relevance of each attribute from the input dataset with respect to the class attribute. In this case the attribute weights are the coefficients of a hyper-plane calculated by a SVM classification model. The third operator is used to select only those attributes that satisfy a pre-specified condition. The condition specified here is "select Top 3 attributes". The process resulted in the following attributes:-

1. X= Attribute [5000 1000 50000] has a weight of 1
2. Y= Attribute [5000 0 500000] has a weight of 0.827
3. Z= Attribute [1000 5000 300000] has a weight of 0.708

Figure (20) shows a 2-D visualization model of AML and ALL cases using attributes X and Z.

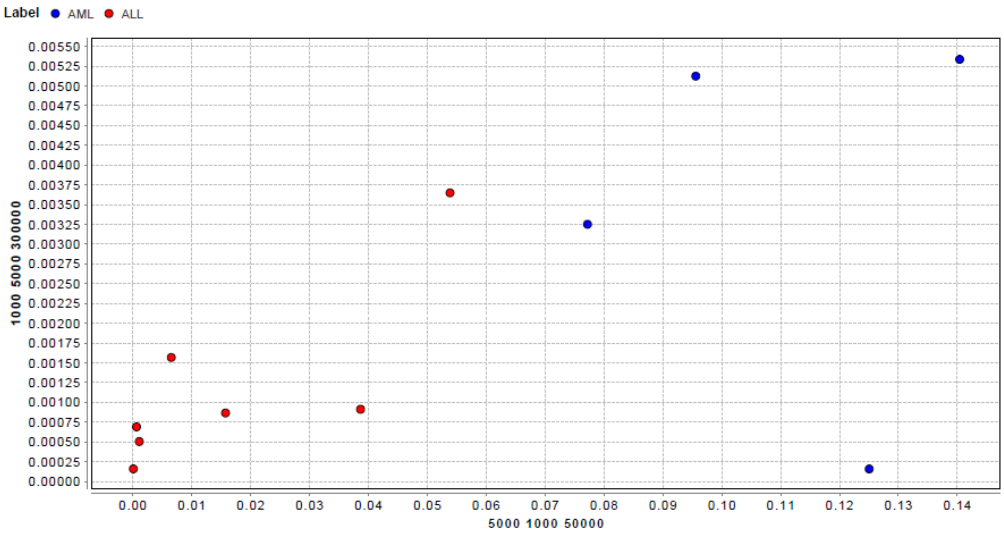


Figure 20: A 2D (X-Z) visualization model for AML and ALL cases

7.5.Analyzing the Reduced Dataset using Rapid Miner 5.3

Table (9) shows the results of analyzing the reduced dataset using Rapid Miner 5.3. Comparing table (8 and 9), it is clear that the performance of all models has enhanced significantly due to using the reduced dataset instead of the (CA) dataset.

#	Model	False AML	False ALL	Accuracy (%)	Precision (%)	Recall (%)
1	Decision Trees	0	0	100	100	100
2	Rule Induction	0	0	100	100	100
3	Logistic Regression	1	2	75	75	85.71
4	K Nearest Neighbor	0	0	100	100	100
5	Perceptron	0	4	60	60	100

6	Naïve Bayes	1	0	95	100	85.71
7	SVM	1	2	75	75	85.71

Table 9: Classification Models for the reduced dataset

Also different clustering techniques succeeded in classifying all cases into (AML) cluster and (ALL) cluster. Table (10) shows the results of clustering the reduced dataset.

#	Clustering Technique	Cluster_0	Cluster_1
1	K-Means	4 AML cases	7 ALL cases
2	Expectation Maximization	3 AML cases	7 ALL cases + 1 AML case
3	K-Means (Kernel)	4 AML cases	7 ALL cases
4	K-Medoids	4 AML cases	7 ALL cases
5	X-Means	4 AML cases	7 ALL cases

Table 10: Results of clustering the reduced dataset

7.6. Classifying (ALL) cases to B-ALL and T-ALL

In order to classify (ALL) cases into either B-ALL or T-ALL; (ALL) cases are extracted from the original (CA) dataset, and are used to form a new (ALL) CA dataset. Figure (21) shows a part of (ALL) CA dataset with seven (ALL) cases. Table (11) shows the results of applying different classification models on (ALL) CA dataset. Classifier models are built using default Rapid Miner 5.3 parameter values, and cross validation (with 10 folds and stratified sampling) is used to measure the performance of the classifier model.

Row No.	Cases	Label	0 0 0	0 0 1000	0 0 10000	0 0 100000	0 0 300000	0 0 5000	0 0 50000	0 0 500000	0 1000 0	0 1000 1000	0 100
1	1B	b	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.001	0.000	0.01
2	1T	t	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0	0.000	0.001	0.00
3	2B	b	0.000	0.002	0.000	0.001	0.000	0.001	0.000	0	0.000	0.001	0.00
4	2T	t	0.000	0.000	0	0.003	0.001	0.000	0.001	0	0.000	0.001	0
5	3B	b	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.001	0.00
6	4B	b	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0	0.000	0.003	0.00
7	4T	t	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.010	0.00

Figure 21: Part of (ALL) Combined Attributes Dataset

#	Model	False B-ALL	False T-ALL	Accuracy (%)	Precision (%)	Recall (%)
1	Decision Trees	0	1	85.71	75	100
2	Rule Induction	0	1	85.71	75	100
3	Logistic Regression	1	1	71.43	66.67	66.67
4	K Nearest Neighbor	1	3	42.86	40	66.67
5	Naïve Bayes	2	1	57.14	50	33.33
6	Support Vector Machines	1	2	57.14	50	66.67

Table 11: Classification Models for (ALL) CA dataset

7.7.Attribute weighting techniques

An attribute weighting process is established in Rapid Miner 5.3 and used to reduce the number of attributes in the (ALL) CA dataset. Using SVM operator, the process resulted in the following top three attributes:-

1. X- Attribute (0 0 300000) with a weight of 1
2. Y- Attribute (5000 5000 100000) with a weight of 0.979.
3. Z- Attribute (0 50000 300000) with a weight of 0.977.

The previous three attributes are used to visualize (ALL) cases and classify them to either B-ALL or T-ALL. Figures (22, 23 and 24) show different visualizations for B-ALL and T-ALL cases.

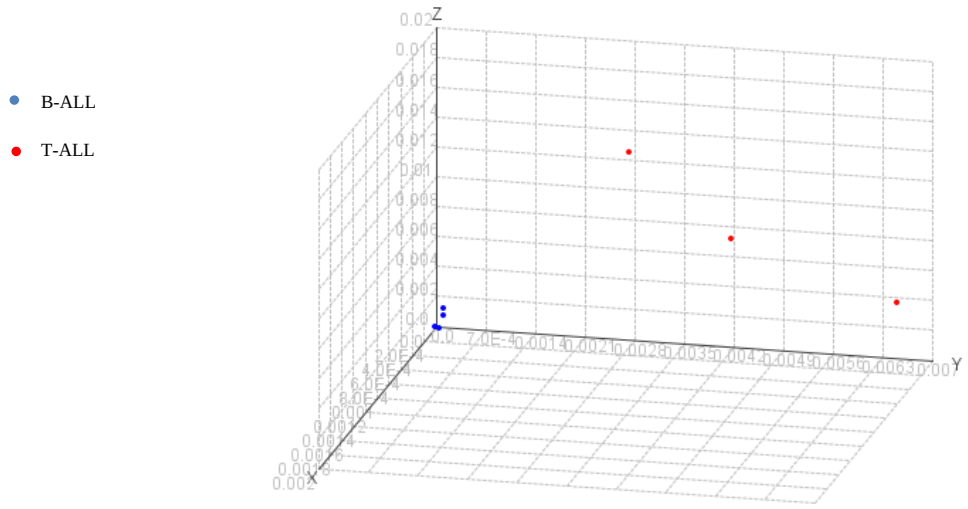


Figure 22: A 3-D (X-Y-Z) visualization model for ALL cases

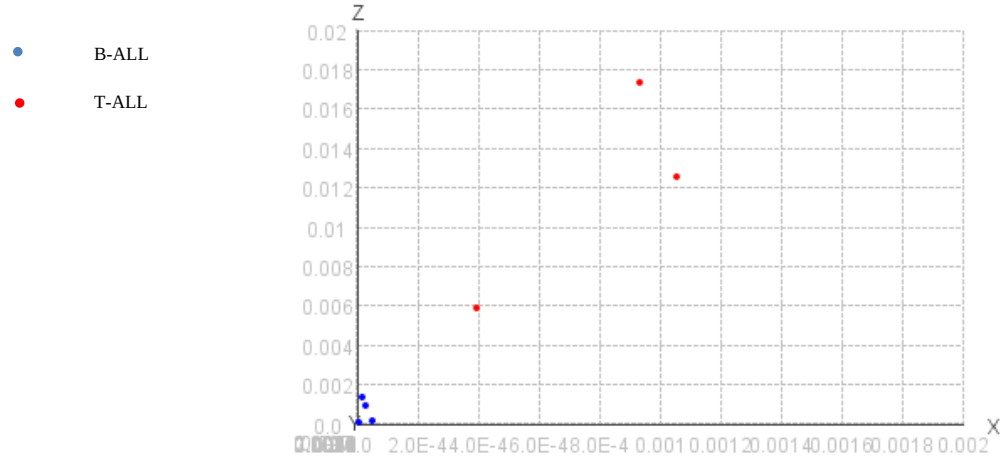


Figure 23: A 2-D (X-Z) scatter Plot of ALL cases

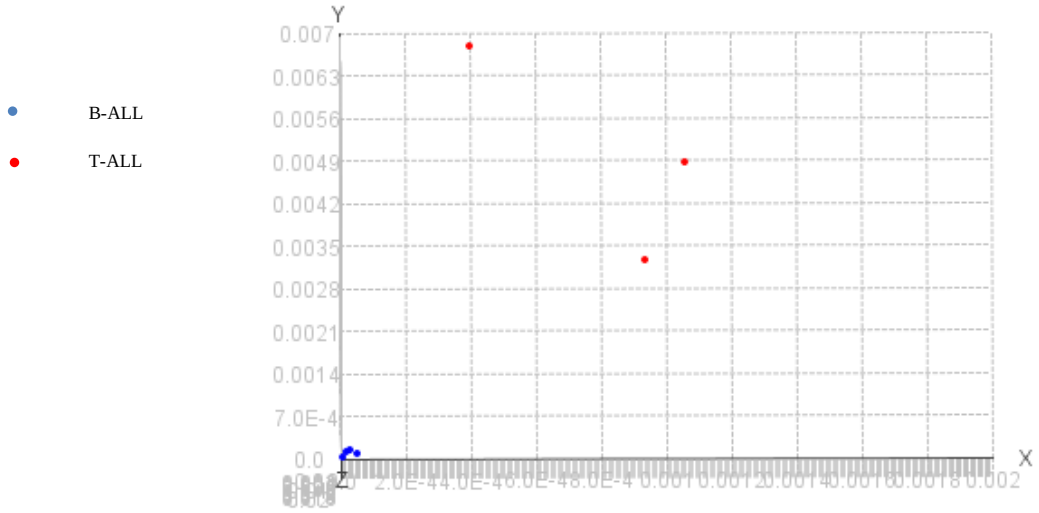


Figure 24: A 2-D (X-Y) scatter Plot of ALL cases

7.8. Classifying the reduced (ALL) dataset

The top three attributes resulted in the previous section can be used in a reduced dataset of seven (ALL) cases. This dataset is classified using different classifier models in Rapid Miner 5.3. Table (12) shows the results of different classifier models for the reduced (ALL) dataset. Comparing tables (11) and (12) verifies the significant enhancement due to using reduced attributes (ALL) dataset.

#	Model	False B-ALL	False T-ALL	Accuracy (%)	Precision (%)	Recall (%)
1	Decision Trees	1	0	85.71	100	66.67
2	Rule Induction	1	0	85.71	100	66.67
3	Logistic Regression	0	0	100	100	100
4	K Nearest Neighbor	0	0	100	100	100
5	Naïve Bayes	0	0	100	100	100
6	SVM	0	0	100	100	100

Table 12: Classification Models for the reduced (ALL) dataset

Chapter 8

8. Conclusion and Future Work

8.1. Summary

Flow cytometry is a very important tool in basic research for investigating many features of cell functions and has emerged as a very important clinical technique in the areas of blood cells and cancer diagnostics. Flow cytometry is used for analyzing a large population of fluorescently labeled cells in a fluid stream. As the particles pass through a focused light source; the amount of light scattered and the emission of a fluorescence label can be measured and stored in a FCS file format.

Conventional analysis method is used in many labs today, which depends mainly on gating. Although gating is an important step in manual analysis techniques; it is a major disadvantage in FCM analysis, as it highly subjective, prone to error and dependent on the experience of the analyst.

The last two decades have seen many developments in computer science, FC industry and instrumentations, biomarkers and fluorochrome conjugate antibodies and the widespread use of immunophenotypic data in diagnosis. All these developments encouraged researches for FCM automated analysis techniques

This thesis developed, presented and tested a framework for representing multi-parameter flow cytometric dataset that contains flow cytometric data of different attributes divided across more than one FCS files. This framework is a novel approach that combines two main concepts: Probability Binning (PB) and Bayesian Inference (BI). Using this combined approach, we could represent FCM data in such a representation that allows us to classify blood flow data to normal and Leukemia incidence cases. Also the same representation could successfully by used to discriminate different Leukemia types (AML, B-ALL or T-ALL) using a two-stage classifier resulting in 100% accuracy and 100% sensitivity for several classification models.

Many researches presented either PB or BI in analyzing flow cytometric data, but not both of them. Researches that applied PB resolved the challenge of unordered cells across different FCS files, but lost the dependencies between features that are divided on multiple FCS files. Other researches applied BI on the individual cell level. These approaches resolved the challenge of

multiple features divided across many FCS files for the same patient, but did not resolve the challenge of different cells' order. Also many researches used partially or fully automated (gating) as a first step in flow cytometric analysis. The approach presented here can work on the raw flow cytometric data without the need for gating.

The currently presented approach assumes that backbone channels are common across all FCS files of the same patient's examinations, and this is practically true. The approach also assumes that supplemental channels are independent of each other given the backbone attribute. This is also true as a solution to flow cytometers' technical limitations; panels with two or more combinations of overlapping antibodies are used to measure cells' characteristics for the complete identification of different hematological malignancies.

A clear limitation of the approach presented here is the explosion of the generated combined attributes. Recall (Equation 13: Number of generated combined attributes) the number of generated combined attributes $N= (b)^{m}$: where (m) is the number of individual attributes, and b is the number of break points used in binning.

This limitation could be overcome by selecting one backbone attribute, which is common across all FCS files (practically, CD45 is the backbone attribute which exist almost in all FCS files) plus two supplemental attributes (which vary according to hematological malignancy under investigation). The selection of supplemental attributes should be done by browsing all FCS files and picking the features that are available for all cases, as the procedure presented here uses the same set of features on all FCS samples. Note that, FCM tests are expensive and practically, not all antibodies and markers are used in FCM tests. For example, the discrimination between AML, B-ALL and T-ALL depends on the set of channels: CD34, CD117, CD19, cyCD3, CD3 and CD45. Practically, CD117, CD19 and cyCD3 are not available for all cases. Thus the procedure was evaluated on only CD45, CD34 and CD3. Also a wise binning sequence must be used by trial and error, taking into account striking a balance between accuracy and compact representation.

The experiments presented in this thesis showed that it is sufficient to depend only on two or three combined attributes (from the global dataset) in order to have a clear and accurate results in either classifying leukemia or in discriminating leukemia types.

8.2.Future Work

It is useful to test the approach presented here in diagnosing different hematological malignancies, with the condition that the approach is applied on the correct set of channels. (Van Dongen et al. 2012) present backbone attributes and supplemental attributes used in the diagnosis of different hematological malignancies. We claim that it is sufficient to depend only on three attributes (one backbone and two supplemental) in our algorithm to get meaningful results with 100% accuracy and sensitivity.

It is also challenging and interesting to test the approach presented here in analyzing and discovering Minimal Residual Diseases (MRD). MRD is the small numbers of cancerous cells in the body's tissues that are not cured during treatment of leukemia. MRD is the most significant cause of recrudescence in leukemia patients following chemotherapy treatment. The algorithm presented here is applied on raw data and gating is not applied in any stage, and thus we claim that this approach can discover MRD even in very low percentages.

Also, as a future extension to this work, it is important to test this approach on different medical research field like DNA analysis and gene selection from microarray data, or in drug discovery platforms.

References

Abcam. (2014). Flow Cytometry Guide. Available form: <http://docs.abcam.com/pdf/protocols/Introduction_to_flow_cytometry_May_10.pdf>. [Accessed on

Aggarwal, C. C. (Ed.). (2014). Data Classification: Algorithms and Applications. CRC Press.

Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., ... & DREAM Consortium. (2013). Critical assessment of automated flow cytometry data analysis techniques. Nature methods, vol. 10, no. 3, pp. 228-238.

Andrade, M. (2010). A note on foundations of probability. Journal of Mathematics and Technology, vol. 1, no. 1, pp. 96-98.

Bashashati, A., & Brinkman, R. R. (2009). A survey of flow cytometry data analysis methods. Advances in bioinformatics, vol. 2009.

Boedigheimer, M. J., & Ferbas, J. (2008). Mixture modeling approach to flow cytometry data. Cytometry Part A, vol. 73, no. 5, pp. 421-429.

Chan, C., Feng, F., Ottinger, J., Foster, D., West, M., & Kepler, T. B. (2008). Statistical mixture modeling for cell subtype identification in flow cytometry. Cytometry Part A, vol. 73, no. 8, pp. 693-701.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), vol. 41, no. 3, pp. 15.

Costa, E. S., Arroyo, M. E., Pedreira, C. E., Garcia-Marcos, M. A., Tabernero, M. D., Almeida, J., & Orfao, A. (2006). A new automated flow cytometry data analysis approach for the diagnostic screening of neoplastic B-cell disorders in peripheral blood samples with absolute lymphocytosis. Leukemia, vol. 20, no. 7, pp. 1221-1230.

Cox, C., Reeder, J. E., Robinson, R. D., Suppes, S. B., & Wheeless, L. L. (1988). Comparison of frequency distributions in flow cytometry. Cytometry, vol. 9, no. 4, pp. 291-298.

Cualing, H. D. (2000). Automated analysis in flow cytometry. Cytometry, vol. 42, no. 2, pp. 110-113.

Duda, R. O., Hart, P. E., & Stork, D. G. (1999). Pattern classification. John Wiley & Sons.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern classification. John Wiley & Sons.

Flow Cytometry: How Does it Work? (2014). Available form: <http://www.unsolvedmysteries.oregonstate.edu/flow_06>. [Accessed on 10 March 2014].

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, vol. 3, pp. 1157-1182.

Lacombe, F., Durrieu, F., Briais, A., Dumain, P., Belloc, F., Bascans, E., ... & Bernard, P. (1997). Flow cytometry CD45 gating for immunophenotyping of acute myeloid leukemia. Leukemia, vol. 11, no. 11, pp. 1878-1886.

Lakoumentas, J., Drakos, J., Karakantza, M., Nikiforidis, G. C., & Sakellaropoulos, G. C. (2009). Bayesian clustering of flow cytometry data for the diagnosis of B-chronic lymphocytic leukemia. Journal of biomedical informatics, vol. 42, no. 2, pp. 251-261.

Lakoumentas, J., Drakos, J., Karakantza, M., Zoumbos, N., Nikiforidis, G., & Sakellaropoulos, G. (2006). The probabilities mixture model for clustering flow-cytometric data: An application to gating lymphocytes in peripheral blood. Biological and Medical Data Analysis, pp. 150-160. Springer Berlin Heidelberg.

Liu, L., Xiong, L., Lu, J. J., Gernert, K. M., & Hertzberg, V. (2008). Comparing and clustering flow cytometry data. In Bioinformatics and Biomedicine, 2008. BIBM'08. IEEE International Conference, pp. 305-309. IEEE.

National Cancer Institute (NCI). (2013). What You Need To Know About Leukemia. Available form: <http://www.cancer.gov/publications/patient-education/wyntk-leukemia>. [Accessed 26 June 2014].

Pedreira, C. E., Costa, E. S., Almeida, J., Fernandez, C., Quijano, S., Flores, J., ... & Orfao, A. (2008). A probabilistic approach for the evaluation of minimal residual disease by multi-parameter flow cytometry in leukemic B-cell chronic lymphoproliferative disorders. Cytometry Part A, vol. 73, no. 12, pp. 1141-1150.

Pedreira, C. E., Costa, E. S., Arroyo, M. E., Almeida, J., & Orfao, A. (2008). A multidimensional classification approach for the automated analysis of flow cytometry data. Biomedical Engineering, IEEE Transactions on, vol. 55, no. 3, pp. 1155-1162.

Pedreira, C. E., Costa, E. S., Barrena, S., Lecrevisse, Q., Almeida, J., van Dongen, J. J., & Orfao, A. (2008). Generation of flow cytometry data files with a potentially infinite number of dimensions. Cytometry Part A, vol. 73, no. 9, pp. 834-846.

Pedreira, C. E., Costa, E. S., Lecrevisse, Q., van Dongen, J. J., & Orfao, A. (2013). Overview of clinical flow cytometry data analysis: recent advances and future challenges. Trends in biotechnology, vol. 31, no. 7, pp. 415-425.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T. I., Maier, L. M., ... & Mesirov, J. P. (2009). Automated high-dimensional flow cytometric data analysis. Proceedings of the National Academy of Sciences, vol. 106, no. 21, p. 8519-8524.

Roederer, M., Moore, W., Treister, A., Hardy, R. R., & Herzenberg, L. A. (2001). Probability binning comparison: a metric for quantitating multivariate distribution differences. Cytometry, vol. 45, no. 1, pp. 47-55.

Roederer, M., Treister, A., Moore, W., & Herzenberg, L. A. (2001). Probability binning comparison: a metric for quantitating univariate distribution differences. Cytometry, vol. 45, no. 1, pp. 37-46.

Rowley, T. (2012). Flow Cytometry - A Survey and the Basics. MATER METHODS, vol. 2, pp. 125. Available from: <http://www.labome.com/method/Flow-Cytometry-A-Survey-and-the-Basics.html>. [Accessed on 20 March 2014].

Select Science. (2014). Flow Cytometry Buying Guide. Available from: <http://www.selectscience.net/downloads/articles/SelectScience_Flow_Cytometry_Buying_Guide_2014.pdf>. [Accessed on 15 June 2014].

Semrock. (2014). Filters for Flow Cytometry. Available form: <http://www.semrock.com/flow-cytometry.aspx>. [Accessed on 12 March 2014].

Shih, M. C., Huang, S. H. S., Donohue, R., Chang, C. C., & Zu, Y. (2013). Automatic B cell lymphoma detection using flow cytometry data. BMC genomics, vol. 14, no. Suppl 7, pp. S1.

Simon, S., Guthke, R., Kamradt, T., & Frey, O. (2012). Multivariate analysis of flow cytometric data using decision trees. Frontiers in microbiology, vol. 3.

Simon, U., Mucha, H. J., & Brüggemann, R. (2005). Model-based cluster analysis applied to flow cytometry data. Innovations in Classification, Data Science, and Information Systems, pp. 69-76. Springer Berlin Heidelberg.

Tadmouri, G. O., Nair, P., Obeid, T., Gallala, S. Genetics Made Easy 2: Cancers, Leukemia, pp. 10-11. Centre for Arab Genomic Studies. Available form: <http://www.cags.org.ae/gme2cancersenleukemia.pdf>. [Accessed on 15 June 2014].

Van Dongen, J. J. M., Lhermitte, L., Böttcher, S., Almeida, J., Van der Velden, V. H. J., Flores-Montero, J., ... & Orfao, A. (2012). EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. Leukemia, vol. 26, no. 9, pp. 1908-1975.

Walther, G., Zimmerman, N., Moore, W., Parks, D., Meehan, S., Belitskaya, I., ... & Herzenberg, L. (2009). Automatic clustering of flow cytometry data with density-based merging. Advances in bioinformatics, vol. 2009.

Zare, H., Shooshtari, P., Gupta, A., & Brinkman, R. R. (2010). Data reduction for spectral clustering to analyze high throughput flow cytometry data. BMC bioinformatics, vol. 11, no. 1, pp. 403.

Appendix A: Applying Probability Binning and Bayesian Inference to represent a sample flow cytometric case as a multi-dimensional histogram of combined attributes

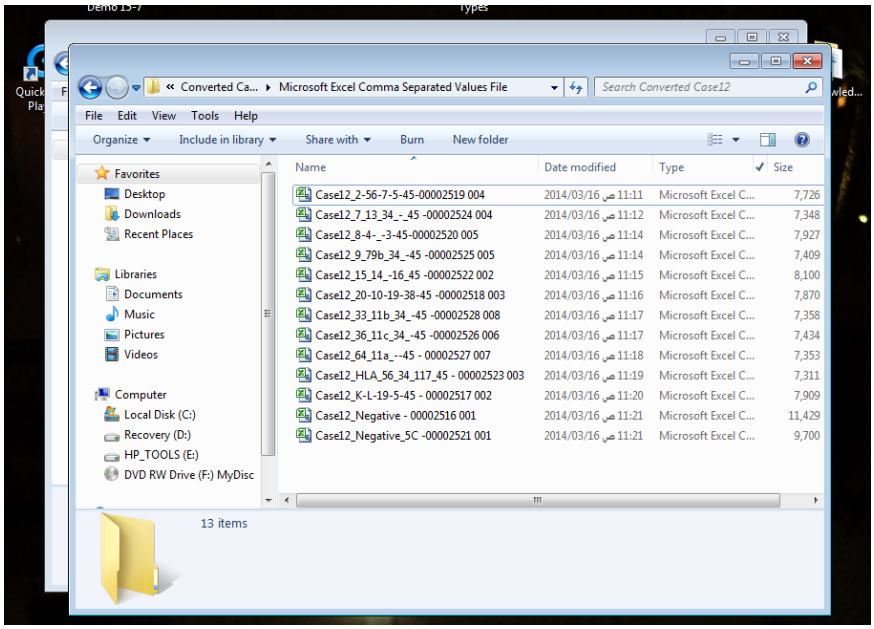


Figure 25: FCS files for one patient (one sample case)

Figure (28) represents different FCS files containing examinations for the same patient. It is clear that CD45 is a common attribute across all files, while CD3 and CD19 are not in the same FCS file.

In order to apply the approach presented here, two FCS files will be selected: Case12_8-4-_-3-45-00002520 005 and Case12_20-10-19-38-45 -00002518 003. The first FCS file contains the attributes CD3 and CD45, while the second FCS file contains the attributes CD19 and CD45.

The following table summarizes the steps of the approach:-

<table>
<tr><th>#</th><th>Step</th><th>Objective</th><th>Result</th></tr>
<tr><td>0</td><td>Load the file “Case12_8-4-_-3-45-00002520 005 .csv” to R</td><td>To read the values of intensities of individual cells in the two channels CD3 and CD45. For example, cell #1 show intensity of (5.6742) in channel CD3, and show intensity of (465.5526) in channel CD45.</td><td>
<table>
<tr><th>Cell #</th><th>CD3</th><th>CD45</th></tr>
<tr><td>1</td><td>5.67422</td><td>465.5526</td></tr>
<tr><td>2</td><td>11.2403</td><td>349.1152</td></tr>
<tr><td>3</td><td>8.13123</td><td>164</td></tr>
<tr><td>4</td><td>2.50286</td><td>2.813318</td></tr>
<tr><td>5</td><td>15.1247</td><td>148.5508</td></tr>
<tr><td>…..</td><td>…..</td><td>…..</td></tr>
<tr><td>63910</td><td>63.2093</td><td>147.2207</td></tr>
</table>
</td></tr>
<tr><td></td><td colspan="3">Stage (A): Applying the modified Probability Binning algorithm
(Convert cells’ intensity distributions to intensity histograms)</td></tr>
<tr><td>1</td><td>Define the possible range of intensities (minimum and maximum intensity values)</td><td>Flow cytometric files containing CD3, CD19 and CD45 are explored to define the minimum and maximum intensity values that can be assigned to a cell.</td><td>Minimum intensity value = [0]
Maximum intensity value =[10000]</td></tr>
<tr><td>2</td><td>The intensity value of each cell is discretized according to the selected bins.</td><td colspan="2">
<table>
<tr><th>Cell #</th><th>Attribute CD3</th></tr>
<tr><td>1</td><td>10</td></tr>
<tr><td>2</td><td>20</td></tr>
<tr><td>3</td><td>10</td></tr>
<tr><td>4</td><td>10</td></tr>
<tr><td>5</td><td>20</td></tr>
<tr><td>….</td><td>……</td></tr>
<tr><td>63910</td><td>70</td></tr>
</table>
</td></tr>
</table>

3	The discretized intensities of individual cells are placed in the appropriate bin, then the number of cells in each bin is counted to form the histogram (probability) at that intensity.	

Bin #	Bin	Cell counts
1	10	24679
2	20	18373
3	30	7016
4	40	3820
5	50	2329
6	60	1640
7	70	1138
8	80	798
9	90	614
10	100	491
11	200	1528
12	300	339
13	400	245
14	500	222
15	1000	498
16	2000	114
17	3000	26
18	4000	14
19	5000	9
20	6000	8
21	7000	3
22	8000	1
23	9000	2
24	10000	2

4	Step (2) is repeated for CD45 in FCS 1	

Cell #	CD45
1	500
2	400
3	200
4	10
5	200
….	
63910	200

5 Repeat step (3) for CD45 in FCS 1

Bin #	Bin	Cell counts
1	10	5540
2	20	769
3	30	385
4	40	228
5	50	204
6	60	166
7	70	157
8	80	245
9	90	417
10	100	714
11	200	23231
12	300	16739
13	400	5822
14	500	2067
15	1000	4958
16	2000	2023
17	3000	166
18	4000	38
19	5000	13
20	6000	10
21	7000	3
22	8000	11
23	9000	3
24	10000	0

6 Stage (B): Generating the joint probability equation of channels (CD3, CD19, CD45)

$$P(CD3, CD19, CD45) = \frac{P(CD3, CD45)P(CD19, CD45)}{P(CD45)} \qquad (2)$$

Stage (C): Generating contingency tables for (CD3, CD45), (CD19, CD45) and frequency table for (CD45)

7 Arrange frequencies of CD3 and CD45 in a contingency table

#	CatCD3	CatCD45	freq
1	10	10	4951
2	10	20	233
3	10	30	94
4	10	40	66
282	10000	4000	1
283	10000	8000	1

8 Pivot Table of (CD3 and CD45)

Sum of freq	CD45																							
CD3	10	20	30	40	50	60	70	80	90	100	200	300	400	500	1000	2000	3000	4000	5000	6000	7000	8000	9000	Grand Total
10	4951	233	94	66	67	56	54	70	130	197	6872	5578	1990	812	2672	832	4	1						24679
20	340	174	46	20	21	16	21	42	96	202	7078	5240	2155	821	1414	643	41	3						18373
30	111	139	48	21	9	15	12	16	41	82	3053	2252	684	165	206	116	40	6						7016
40	54	96	36	13	15	7	7	11	18	47	1710	1236	366	96	55	32	16	4	1					3820
50	15	51	41	16	11	8	6	12	19	26	1136	722	186	45	23	8	2	1	1					2329
60	15	28	45	14	12	9	4	4	14	16	786	507	136	27	15	6	2							1640
70	9	11	30	11	7	6	4	12	6	12	585	324	91	16	12	1		1						1138
80	7	14	21	13	8	5	5	7	10	16	411	217	46	6	9	2	1							798
90	8	6	6	17	14	2	4	6	8	24	312	157	32	8	10									614
100	4	3	8	9	6	1	1	6	9	12	266	130	25	6	2	2		1						491
200	14	11	9	24	30	33	27	37	40	59	813	288	68	30	40	5								1528
300	3	1	1	2	2	4	11	15	22	15	131	36	14	10	56	16								339
400	3	1		1	2	3	1	6	2	4	48	13	7	7	112	33		2						245
500						1		1	2	1	18	11	3	2	113	69	1							222
1000	1	1		1						1	12	27	17	9	193	216	20							498
2000	1											1	2	7	22	37	31	11	2					114
3000	1														4		7	6	5	3				26
4000	2															3			3	4	2			14
5000																2				2		3	2	9
6000	1																1	1	1	1	1	1	1	8
7000																						3		3
8000																						1		1
9000																						2		2
10000																		1				1		2
Grand Total	5540	769	385	228	204	166	157	245	417	714	23231	16739	5822	2067	4958	2023	166	38	13	10	3	11	3	63909

9 steps (1 to 6) are repeated for both CD19 and CD45 in the second FCS file

	CatCD19	CatCD45	freq
1	10	10	3941
2	10	20	551
3	10	30	213
4	10	40	151
	….		
238	6000	8000	1
239	6000	9000	7

10 Pivot table of (CD19 and CD45)

Sum of freq	Column Labels																								
Row Labels	10	20	30	40	50	60	70	80	90	100	200	300	400	500	1000	2000	3000	4000	5000	6000	7000	8000	9000	10
10	3941	551	213	151	103	86	75	74	112	222	9918	9180	3368	1206	3347	1511	44	1						
20	88	25	64	73	53	33	36	45	93	180	8690	5780	1878	666	878	320	69	4	1					
30	36	11	5	5	7	16	13	23	42	65	2276	1502	376	113	151	23	17	7						
40	25	13	3	1	4	4	9	15	18	39	1028	568	141	27	35	7	4		1					
50	10	7	3	1		3	5	4	16	26	498	231	43	19	10	3	5	1						
60	10	4	1	1	1	1	7	4	8	14	269	147	31	8	17	2	2							
70	6	2	2				2	3	6	8	152	84	17	8	13	3								
80	2	2		1	2	1	2	2	7	8	83	46	8	7	14	4		1						
90	2	1	1			1	1	4	3	3	52	28	4	4	17		2							
100	1	2	2	1			1	3	1	6	38	18	5	9	26	3								
200	6	4	4	2	2	2	3	6	7	5	79	138	94	57	354	105	2							
300	2		1		1	2				1	5	30	30	24	197	100	2							
400		1								1	2	4	6	8	46	51	4							
500											1		1	1	9	27	11			1				
1000						1					5	4		1	3	21	39	21	7					
2000				1													3	3	5	10	3		1	
3000																					1		2	
4000																						1	5	
5000																							2	
6000	1														1		1		1		1	1	7	
Grand Total	4130	623	299	237	173	150	154	183	313	578	23096	17760	6002	2158	5118	2180	205	38	15	11	5	2	17	

11	Calculate frequency table for CD45 as a univariate data.	

CD45	Freq	P
10	**5540**	0.086686
20	**769**	0.012033
30	**385**	0.006024
40	**228**	0.003568
50	**204**	0.003192
60	**166**	0.002597
70	**157**	0.002457
80	**245**	0.003834
90	**417**	0.006525
100	**714**	0.011172
200	**23231**	0.363501
300	**16739**	0.261919
400	**5822**	0.091098
500	**2067**	0.032343
1000	**4958**	0.077579
2000	**2023**	0.031654
3000	**166**	0.002597
4000	**38**	0.000595
5000	**13**	0.000203
6000	**10**	0.000156
7000	**3**	4.69E-05
8000	**11**	0.000172
9000	**3**	4.69E-05

12	Apply equation (12)	

#	Pattern	Pjoint
1	10 10 10	0.05551
2	10 10 20	0.002631
3	10 10 30	0.00082
4	10 10 40	0.000689
5	10 10 50	0.000533
6	10 10 60	0.000457
7	10 10 70	0.000407
8	10 10 80	0.000333
…..	……	…….
11089	8000 7000 10000	0