Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Seminarausarbeitung

Jan-Christoph Meier

Anwendung von Data Mining auf Daten der Durchflusszytometrie

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Inhaltsverzeichnis

1 Einführung

Mithilfe der Durchflusszytometrie wird in der Multiple Sklerose Forschung am Zentrum für molekulare Neurobiologie Hamburg (ZMNH) das Blut von Multiple Sklerose erkrankten Personen untersucht. Ziel hierbei ist es, die im Blut vorhandenen Zellpopulationen zu bestimmen, was Aufschluss über den aktuellen Stand der Krankheit geben kann.
Das Verfahren wird auch als »Fluorescence activated cell sorting« (FACS) und die dabei entstehenden Daten, als FACS-Daten bezeichnet.
Diese FACS-Daten wurden über mehrere Jahre im Rahmen verschiedener Studien erhoben und in einer sogenannten Biobank gespeichert. Diese enthält Daten von mehreren hundert Patienten und tausenden von Proben. Zusätzlich wurden noch weitere klinische Daten, beispielsweise über Krankheitsverläufe und eingenommene Medikamente erhoben und gespeichert.

1.1 Motivation

Die in der Biobank vorhandenen Daten haben einen hohen wissenschaftlichen Wert, da eine derart umfangreiche Erfassung von erkrankten Personen bisher nur selten durchgeführt wurde. Von besonderem Interesse ist die Koppelung der Daten aus der Durchflusszytometrie (FACS-Daten) mit den klinischen Daten, wobei eine Korrelation zwischen klinischen Parametern, wie zum Beispiel Krankheitsverlauf und den Zellpopulationen im Blut, hergestellt werden kann.
Eine umfangreiche Analyse des Datenbestandes ist zum aktuellen Zeitpunkt mit einem sehr hohen Aufwand verbunden. Vor der Analyse müssen die Daten vorverarbeitet werden, um bestimmte Zellpopulationen, die von besonderem Interesse sind, aus den Daten zu extrahieren. Dies erfolgt zurzeit mit der Software FACSDiva, die keine Möglichkeit bietet, die Verarbeitung zu automatisieren. Ein weiteres Problem stellt die unzureichende Organisation der Daten dar. Diese sind sehr willkürlich in Ordnerstrukturen im Dateisystem organisiert, was das Finden von Daten anhand bestimmter Kriterien so gut wie unmöglich macht.
Eine möglichst automatisierte Analyse des gesamten Datenbestandes, der sich mitlerweile in einer Größenordnung von mehreren Terabyte bewegt, ist sehr wünschenswert.

1.2 Einführung in die Durchflusszytometrie

Mit der Durchflusszytometrie (vgl. Perfetto u. a. (2004)) können Zellen im Blut gemessen werden. Hierfür wird das Blut durch eine dünne Messkammer geleitet und mit einem Laser beschossen, wobei die Signale des Lasers von zwei Sensoren detektiert werden. Diese können sowohl das Seitwärts- als auch das Vorwärts-Streulicht bestimmen. Das Vorwärts-Streulicht gibt Auskunft über die Größe der Zelle, das Seitwärts-Streulicht über die Körnigkeit (Granularität) der Zelle. In der Abbildung 1.1 ist die Anwendung der Durchflusszytometrie schematisch dargestellt.
Neben diesen zwei Detektoren gibt es noch weitere Sensoren für farbiges Licht. Möchte man

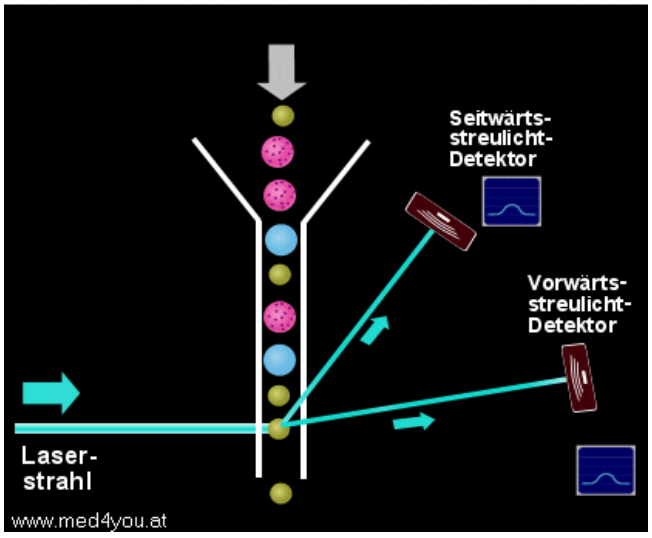

Abbildung 1.1: Schematik der Durchflusszytometrie (Quelle: http://med4you.at)

die Häufigkeiten bestimmter Zelltypen bestimmen, zum Beispiel die von T-Lymphozyten, die eine Untergruppe der Lymphozyten sind, wird ein eingefärbter Antikörper zum Blut hinzugefügt. Dieser Antikörper reagiert auf bestimmte Zelltypen und haftet an ihrerer Oberfläche. Durch die Einfärbung erzeugt der Laser farbiges Licht, sobald er auf die Zelle trifft. Diese farbigen Signale werden dann gemessen, um die Häufigkeiten der unterschiedlichen Zelltypen zu bestimmen.
Bei der Analyse einer Blutprobe werden mehrere hundertausend Signale des Lasers erfasst. Hierbei besteht ein Ereignis des Lasers aus den Signalen der Detektoren für das Seitwärts- und Vorwärts-Streulicht, sowie des farbigen Laserlichts. Zur Visualisierung der Messung werden die Signale in ein X-Y-Diagramm aufgeführt, wobei die X- und Y-Achse jeweils einen Sensor darstellen. In der Abbildung 1.2 ist das Messergebnis für eine Blutprobe dargestellt, hierbei wurden die Signale des Vorwärts- und des Seitwärts-Streulichtes im Diagramm auf-

getragen. Jeder Punkt stellt ein Ereignis des Lasers dar.

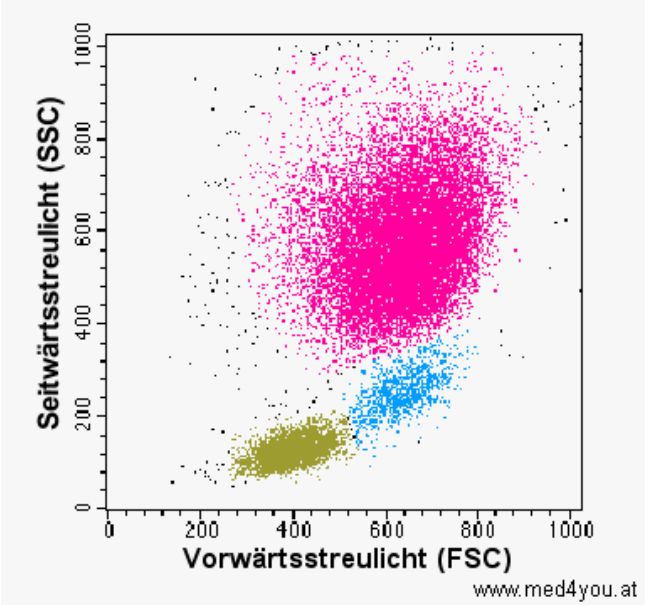

Abbildung 1.2: Messergebnis für eine Blutprobe (Quelle http://www.med4you.at)

Anhand der Signale können die Zelltypen im Blut unterschieden werden. So haben Lymphozyten ein geringes Vorwärts-Streulicht, das heißt eine geringe Größe und ein geringes Seitwärts-Streulicht und somit eine geringe Granularität. In der Abbildung 1.3 ist veranschaulicht, wie die verschiedenen Zelltypen anhand der Signale bestimmt werden können. Bei der Diagnose des aktuellen Status einer Erkrankung werden die Häufigkeiten bestimmter Zellen im Blut ermittelt. Der Grund für eine geringe Anzahl an Lymphozyten kann zum Beispiel eine AIDS-Erkrankung oder Autoimmunerkrankung wie Multiple Sklerose sein.

Gating

Das Gating stellt einen elementaren Schritt bei der Analyse der FACS-Daten dar. Hierbei werden in mehreren Schritten Signale des Lasers selektiert, um Häufigkeiten verschiedener Zellen bestimmen zu können.
In Abbildung 1.4 ist das Gating veranschaulicht. Mit der Selektion R1 wurden ausschließlich Lymphozyten selektiert. Die hierbei ausgewählten Signale des Lasers werden dann in ein weiteres Diagramm überführt. Hierbei werden die farbigen Signale gegenübergestellt. In der Abbildung 1.4 wurden die Kanäle CD19-PE und CD3-FITC aufgetragen. Der Kanal CD19-PE misst Signale von Antikörpern, die auf B-Zellen haften, der Kanal CD3-FITC die Signale von Antikörpern, die auf T-Zellen haften. Da es sich bei T-Zellen sowie B-Zellen um eine Untergruppe der Lymphozyten handelt, wurden im ersten Schritt des Gatings ausschließlich

Lymphozyten selektiert.
Die Selektion beim Gating kann nicht immer eindeutig durchgeführt werden, da teilweise Messfehler in den Daten vorhanden sind.

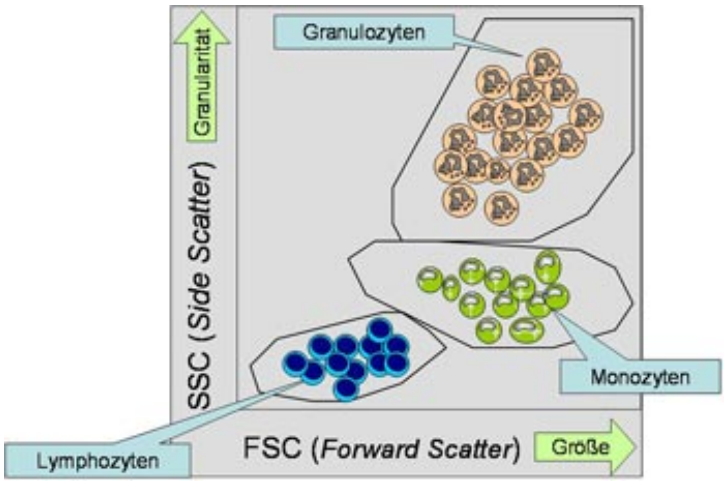

Abbildung 1.3: X-Y-Dot-Plot der Signale einer Messung (Quelle http://www.med4you.at)

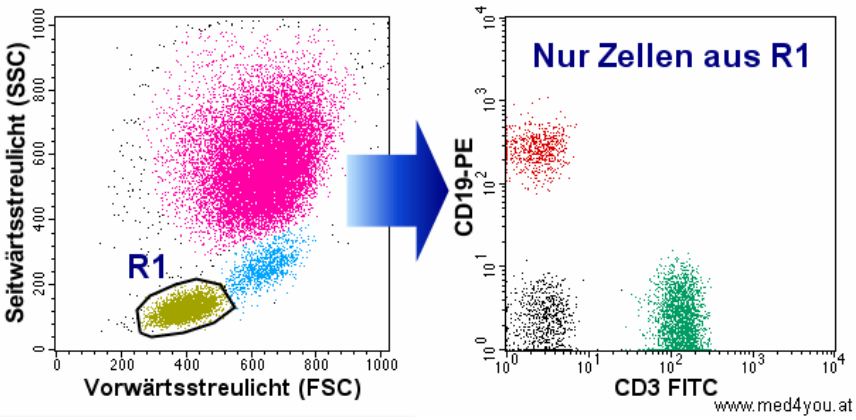

Abbildung 1.4: Selektion von Lymphozyten (Quelle http://www.med4you.at)

2 Masterarbeit

Die Masterarbeit lässt sich in insgesamt drei Teilziele aufteilen. Im ersten Schritt wird die Organisation der Daten durch Entwicklung eines Data Management System verbessert. Im zweiten Schritt wird ein Verfahren entwickelt, das bisher manuell durchgeführte Gating der Daten zu automatisieren. Im finalen Schritt kann Data Mining auf den Datenbestand angewendet werden.

2.1 Zielsetzung

Die FACS-Daten sind zurzeit in einfachen Verzeichnisstrukturen im Dateisystem organisiert. Diese wurden individuell von der jeweiligen Person, die die Daten erfasst hat, angelegt und sind größtenteils sehr inkonsistent benannt. Damit komplexe Analysen der Daten durchgeführt werden können, ist es wünschenswert, diese anhand verschiedener Kriterien selektierbar zu machen. Das kann zum Beispiel anhand der analysierten Zelltypen, des Patienten oder einer bestimmten Studie erfolgen. Die aktuelle Datenorganisation bietet hierfür kaum Möglichkeiten, da die Daten ohne Erfassung weiterer Meta-Informationen abgespeichert werden.
Das Gating der Daten wird mit der Software FACSDiva[1] durchgeführt, die keine Möglichkeit zur Automatisierung bietet. Die Selektion der Zellen beim Gating erfolgt manuell und basierend auf Erfahrung oder »scharfes Hinsehen« der jeweiligen Person, die eine Analyse durchführt. Diese sehr aufwändige manuelle Vorverarbeitung soll mithilfe von Algorithmen automatisiert werden und ohne Interaktion mit dem Benutzer durchgeführt werden.
Im dritten Schritt werden die Daten aus der Durchflusszytometrie mit klinischen Forschungsdaten zusammengeführt. Diese beinhalten unter anderem Krankheitsverläufe und eingenommene Medikamente der Patienten, deren Blut untersucht wurde. Mithilfe von Data Mining Algorithmen soll eine Korrelation zwischen den in den Blutproben vorhandenen Zellpopulationen und den, in der klinischen Datenbank erfassten, Informationen über Krankheitsverlauf und eingenommene Medikamente hergestellt werden.

[1] http://www.bdbiosciences.com/instruments/software/facsdiva/index.jsp

2.2 Vorgehen

Die Rohdaten aus der Durchflusszytometrie liegen im sogenannten »FACS«-Dateiformat vor. Hierbei wird die Analyse einer einzelnen Blutprobe jeweils als eigenständige Datei abgespeichert. Zusätzlich werden Metadaten, die unter anderem das Datum der Analyse und die gemessenen Kanäle enthalten, mit in der Datei hinterlegt. Die Messergebnisse des Lasers sind als Matrix abgespeichert, wobei jede Zeile ein Ereignis des Lasers, also die Erfassung einer Zelle, darstellt. In der Tabelle 2.2 ist die Matrix beispielhaft aufgeführt, die Signale werden als einfache numerische Werte gespeichert, wobei ein hoher Wert ein starkes Signal des Lasers darstellt.

Zeitpunkt	SSC-A	FSC-A	PACIFIC-BLUE
1	5324	330	10254
2	150	730	4504
3	4234	6882	9253
..	..	..	..

Tabelle 2.1: FACS-Datensatz

In der Bioinformatik ist die Programmiersprache R[2] sehr verbreitet. Zur Verarbeitung von Daten aus der Durchflusszytometrie stehen verschiedene Bibliotheken zur Auswahl, diese werden in dem Bioconductor-Projekt[3] organisiert. Mit der Bibliothek »flowCore« (vgl. Hahne u. a. (2009)) kann direkt auf die FACS-Daten zugegriffen werden und sowohl die Signale des Lasers als auch die Metainformationen extrahiert werden.
Desweiteren bietet sich R für die Datenanalyse sehr an, da eine umfangreiche Sammlung an Methoden und Bibliotheken für mathematische Berechnungen und Statistik enthalten sind.

Entwicklung eines Data Management Systems

Die Organisation des Datenbestandes soll durch Entwicklung eines Data Management Systems (DMS) wesentlich verbessert werden. Dieses soll über eine Suchfunktion verfügen, um die Daten anhand verschiedener Parameter zu finden. Das DMS basiert auf einer relationalen Datenbank, die zum Beispiel MySQL sein kann. Aus den FACS-Rohdaten werden die Metadaten extrahiert und in die Datenbank übertragen. Über eine Weboberfläche kann auf die Daten zugegriffen und automatisierte Analysen gestartet werden.

[2] http://www.r-project.org
[3] http://www.bioconductor.org

Automatisieren des Gatings

Es stehen verschiedene Verfahren und Ansätze zur Auswahl, um ein automatisiertes Gating durchzuführen (vgl. Bashashati und Brinkman (2009)). In der Masterarbeit werden Clustering-Algorithmen eingesetzt, um die Selektion der Zellen beim Gating vorzunehmen. Für die Programmiersprache R stehen mit flowMeans (vgl. Nima Aghaeepour und Brinkman (2011)), flowClust (vgl. Lo u. a. (2009)) und flowPeaks (vgl. Ge und Sealfon (2012)) drei Implementierungen von Clustering-Algorithmen zur Verfügung, die auf FACS-Daten angewendet werden können.
Sofern sich die bereits zur Verfügung stehenden Verfahren als nicht optimal für den Datenbestand erweisen, ist es denkbar, dass ein eigenes Verfahren zur Durchführung des Gatings entwickelt wird.

Analyse mit Data Mining

Sobald es möglich ist das Gating automatisiert durchzuführen, können die Zellpopulationen der Daten untersucht werden. Hierfür müssen diese mit den klinischen Daten zusammengeführt werden.
Eine Möglichkeit für die Analyse ist es, einen Zusammenhang zwischen den im Blut vorhandenen Zellen und den eingenommen Medikamenten herzustellen. So kann eine Schlussfolgerung getätigt werden, ob ein Medikament wirkt. In der Tabelle 2.2 ist eine Liste von Beispieldatensätzen aufgeführt. Hierbei wurden die Zellpopulation, das eingenommene Medikament und der Krankheitszustand gegenübergestellt.
Aus dieser vereinfachten Darstellung könnte abgeleitet werden, dass das Medikament M2 nicht wirkt. Diese Schlussfolgerungen könnten anhand einer Vielzahl von Daten, mithilfe von Algorithmen aus dem Bereich »Frequent pattern mining« (vgl. Han u. a. (2000)), erzeugt werden.

Zellpopulation	Medikament	Krankheitszustand
Wenig Lymphozyten	Keine Medikamente eingenommen	Schlecht
Viele Lymphozyten	M1	Gut
Wenig Lymphozyten	M2	Schlecht
..	..	..

Tabelle 2.2: Klinische- und FACS-Daten zusammengeführt

Desweiteren ist es denkbar, dass versucht wird, mithilfe von Algorithmen langfristige Veränderungen in den Datensätzen zu analysieren. Beim »Frequent pattern mining« wird nur ein aktueller Zustand analysiert, jedoch keine langfristige Veränderung. Interessant wäre es zu

analysieren, wie sich der Gesundheitstatus von Patienten, zum Beispiel bei Einnahme verschiedener Medikamente, über einen Zeitraum verändert.
In Projekt 1 und Projekt 2 wurde eine Plattform entwickelt, die verschiedene Data Mining Algorithmen zur Verfügung stellt und für die Analyse der FACS-Daten verwendet werden kann. In Abbildung 2.1 ist die schematische Architektur der Plattform aufgeführt. Die Daten werden als JSON-Datenobjekt über einen Webservice an die Plattform gesendet. Diese kann dann verschiedene Algorithmen auf den Daten ausführen und gibt das Ergebnis als JSON-Datenobjekt zurück.

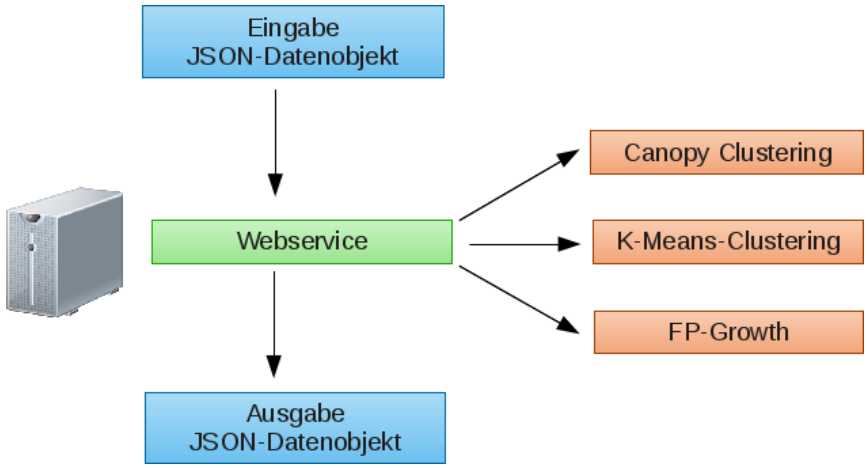

Abbildung 2.1: Data Mining Plattform

Entwicklung der Analyseplattform

In der Masterarbeit soll eine Analyseplattform entstehen, die es ermöglicht die FACS-Daten in mehreren Schritten zu analysieren. Die grundlegende Aufbau dieser Plattform ist in Abbildung 2.2 dargestellt.
Bevor die Benutzer mit den Daten arbeiten können müssen diese in das »Data Management System« import werden. Der Benutzer kann dann eine Selektion, anhand verschiedener Kriterien, auf den Daten durchführen. Diese Kriterien könnten zum Beispiel sein, dass es sich um einen bestimmten Patient oder eine Studie handeln soll. Für die selektierten Daten wird dann mit einem automatisierten Verfahren das Gating durchgeführt. Die daraus resultierenden Zellpopulationen werden an den Webservice der Data Mining Plattform gesendet.
Im zweiten Schritt führt der Benutzer eine Selektion von klinischen Daten durch, die den zuvor selektierten FACS-Daten zugeordnet sind. Diese klinischen Daten werden ebenfalls an den Webservice für die Analyse gesendet.
Der Webservice liefert ein Ergebnis, das dem Benutzer präsentiert wird.

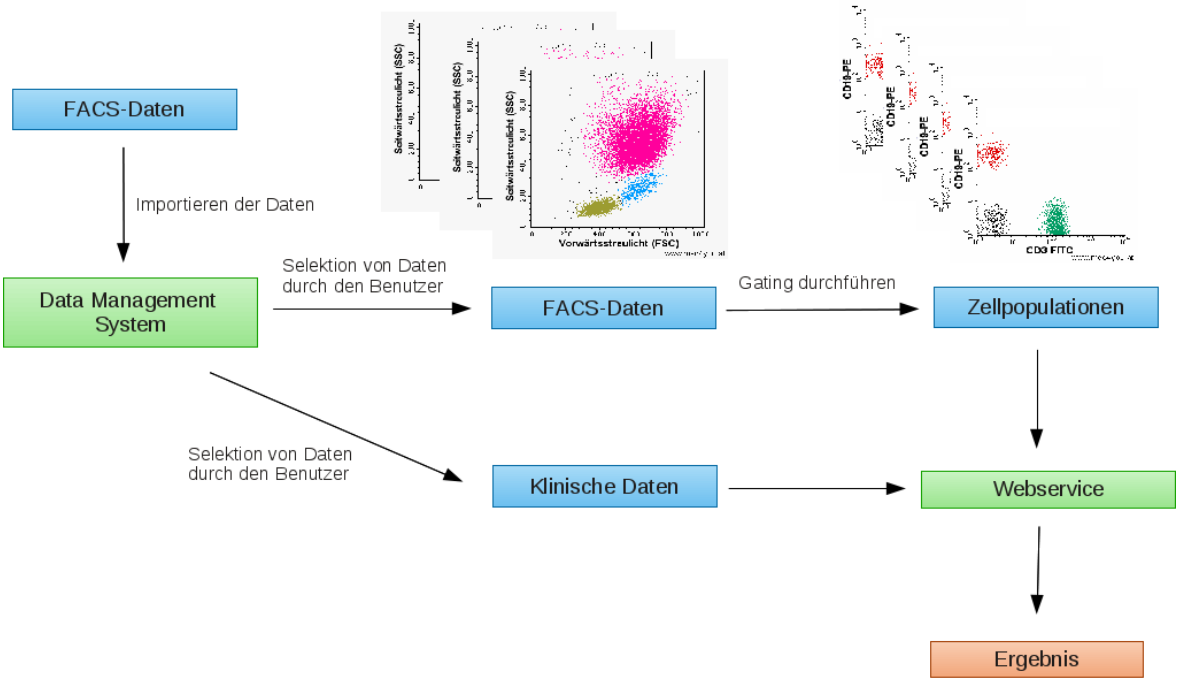

Abbildung 2.2: Schematik der Analysplattform

2.3 Chancen und Risiken

Das in der Masterarbeit zu entwickelnde System ermöglicht es den großen Datenbestand an FACS-Daten zu analysieren und so neue Erkenntnisse für die Multiple Sklerose Forschung zu gewinnen. Die Daten wurden erfasst, konnten jedoch bisher nur mit hohem Aufwand ausgewertet werden. Die Zeitnahe Analyse des gesamten Datenbestandes ist von großem Interesse.
Das Projekt birgt allerdings ein paar Risiken. So kann es passieren, dass als Ergebnis der Analyse nur ein »weißes Rauschen« geliefert wird, was keine Aussagekraft hat. Weiterhin ist es möglich, dass die Daten in einem so inkonsisten Zustand sind, dass es nicht möglich ist eine Analyse über den gesamten Datenbestand durchzuführen. Ein weiteres Risiko ist, das die automatisierte Durchführung des Gatings nur auf einen sehr kleinen Teil der Daten anwendbar ist und so wiederum doch viel Zeit in die Vorverarbeitung der Daten investiert werden müsste.

3 Fazit und Ausblick

Durch das in der Masterarbeit zu entwickelnde System wird es möglich, völlig neue Erkenntnisse aus den Daten der Durchflusszytometrie zu gewinnen. Die Daten wurden bisher nur in aufwändigen manuellen Analysen untersucht und eine umfangreiche Analyse des gesamten Datenbestandes war nicht möglich. Mithilfe verschiedener Data Mining Algorithmen kann ermöglicht werden, neues Wissen aus den Daten zu gewinnen, wobei besonders das Zusammenführen von klinischen Daten und Daten aus der Durchflusszytometrie vielversprechend ist.
Sollte sich das entwickelte System bewähren, ist es denkbar, dass es noch in weiteren Instituten eingesetzt wird, in denen FACS-Daten erfasst werden.
Im Rahmen der Masterarbeit werden Clustering-Algorithmen verwendet, um das Gating der Daten zu automatisieren. Hier ist der Einsatz weiterer Technologien aus der künstlichen Intelligenz, wie zum Beispiel Neuronale-Netze oder »Support Vector Machines« denkbar.

Literaturverzeichnis

[Bashashati und Brinkman 2009] BASHASHATI, Ali ; BRINKMAN, Ryan R.: A Survey of Flow Cytometry Data Analysis Methods. In: *Adv. Bioinformatics* 2009 (2009)

[Ge und Sealfon 2012] GE, Y. ; SEALFON, S.C.: flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. In: *Bioinformatics* 28 (2012), Nr. 15, S. 2052–8

[Hahne u. a. 2009] HAHNE, Florian ; LEMEUR, Nolwenn ; BRINKMAN, Ryan R. ; ELLIS, Byron ; HAALAND, Perry ; SARKAR, Deepayan ; SPIDLEN, Josef ; STRAIN, Errol ; GENTLEMAN, Robert: flowCore: a Bioconductor package for high throughput flow cytometry. In: *BMC Bioinformatics* 10 (2009), S. 106

[Han u. a. 2000] HAN, Jiawei ; PEI, Jian ; YIN, Yiwen: Mining frequent patterns without candidate generation. In: *SIGMOD Rec.* 29 (2000), Mai, Nr. 2, S. 1–12. – ISSN 0163-5808

[Lo u. a. 2009] LO, Kenneth ; HAHNE, Florian ; BRINKMAN, Ryan R. ; GOTTARDO, Raphael: flowClust: a Bioconductor package for automated gating of flow cytometry data. In: *BMC Bioinformatics* 10 (2009), Nr. 1, S. 145

[Nima Aghaeepour und Brinkman 2011] NIMA AGHAEEPOUR, Holger H. H. ; BRINKMAN, Ryan R.: Rapid cell population identication in ow cytometry dat. In: *Cytometry Part A* 79A (2011), Januar, S. 6–13

[Perfetto u. a. 2004] PERFETTO, S. P. ; CHATTOPADHYAY, P. K. ; ROEDERER, M.: Seventeen-colour flow cytometry: unravelling the immune system. In: *Nat Rev Immunol* 4 (2004), August, Nr. 8, S. 648–655. – ISSN 1474-1733