CZECH TECHNICAL UNIVERSITY IN PRAGUE

Faculty of Information Technology

Department of Theoretical Computer Science

Master's Thesis

Supporting the Diagnosis of Borreliosis by Machine Learning Methods

Bc. Jan Motl

Supervisor: Ing. Pavel Kordík, Ph.D.

8th March 2013

Acknowledgements

I would like to thank my family and friends
for support during writing this thesis.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended. In accordance with Article 46(6) of the Act, I hereby grant a nonexclusive authorization (license) to utilize this thesis, including any and all computer programs incorporated therein or attached thereto and all corresponding documentation (hereinafter collectively referred to as the "Work"), to any and all persons that wish to utilize the Work. Such persons are entitled to use the Work in any way (including for-profit purposes) that does not detract from its value. This authorization is not limited in terms of time, location and quantity. However, all persons that makes use of the above license shall be obliged to grant a license at least in the same scope as defined above with respect to each and every work that is created (wholly or in part) based on the Work, by modifying the Work, by combining the Work with another work, by including the Work in a collection of works or by adapting the Work (including translation), and at the same time make available the source code of such work at least in a way and scope that are comparable to the way and scope in which the source code of the Work is made available.

In Prague on 8th March 2013 . . . . . .

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis

Motl, Jan. Supporting the Diagnosis of Borreliosis by Machine Learning Methods. Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2013.

Abstract

Each year over two million people are infected with Lyme disease, the most common tick-borne infection. And while this disease is not commonly lethal it can lead to disability if not treated in time. Unfortunately there is not any reliable diagnosis test to identify Lyme disease. One of such laboratory tests is called Western blot. And even thought Western blots were evaluated many times with statistical methods, there is not a single article about evaluating Lyme disease Western blots with machine learning methods beyond clustering individual blots.

Hence the task of this thesis is to evaluate possibilities of improving prediction accuracy of Lyme disease Western blots with machine learning methods. To assess the possibilities 4227 Western blots were collected and analyzed. Up to the date this is the biggest amount of Lyme disease Western blots ever evaluated in a single study.

In SOM analysis two groups of patients, Lyme positive and negative, naturally separated and Jarisch-Herxheimer reaction was observed.

Keywords: Lyme disease, Borrelia, Western blot, machine learning.

Abstrakt

Každý rok přes dva miliony lidí onemocní nejběžnější nemocí přenášenou klíšťaty, Lymskou boreliózou. A i když není tahle nemoc obvykle smrtelná, může zanechat trvalé následky, není-li diagnostikovaná a léčena včas. Bohužel ale v současnosti neexistuje jediná spolehlivá metoda diagnózy Lymské boreliózy. Jednou takovou laboratorní metodou je Western blot. A ačkoliv Western bloty byly mnohokrát zkoumány pomocí statistických metod, nebyl nalezen článek zkoumající Western bloty metodami strojového učení, pomineme-li hierarchické shlukování.

Tahle práce si klade za úkol prozkoumat možnosti aplikovatelnosti samoorganizačních map na Western bloty. Samoorganizační mapa byla vyhodnocena na vzorku 4227 Western blotů, což je zatím největší množství blotů, které kdy bylo v jediné studii o Lymské borelióze analyzováno.

V mapě se separovali shluky trpící Lymskou boreliózou a pacienti bez Lymské boreliózy. Dále byla v mapě identifikována Jarisch-Herxheimerova reakce.

Klíčová slova: Lymská borelióza, Western blot, strojové učení.

Contents

1 Introduction

1.1 Lyme disease

Lyme disease is a tick-borne illness caused by bacteria Borrelia, nicknamed "the big imitator". Borrelia got its nickname for its high variability. And this variability causes mistakes in diagnosis and treatment (Garaiová, 1990).

High Borrelia variability is caused by its unique genome composing from a linear chromosome and at least 12 linear and 9 circular plasmids. No other bacteria are known to have such high amount of plasmids in a cell (Brorson, 2009).

There are currently 13 known strains of Borrelia and each strain causes different symptoms. For example, *Borrelia hermsii* and *Borrelia parkeri* cause a relapsing fever whereas *Borrelia garinii* is responsible for causing neurological problems. Originally each strain had a different habitat; however, nowadays all 13 strains are present in central Europe (Franke, 2013). In the Czech republic around 4000 patients are diagnosed with Lyme disease yearly (Valtameri, 2011).

1.2 Diagnosis

Because Borrelia is so variable, there is not a single reliable diagnosis test. Lyme disease tests can be divided into three groups: direct microbiological observation (culture), measure of immune response to the infection (Western blot, ELISA) or detection of foreign genetic material (PCR).

Culture is the "gold standard" test for identifying bacteria. A sample of the organism is taken from the patient, is allowed to grow in a medium and then identified. While culture is used to diagnose many infections, it is not practical for Lyme. It's because Borrelia grow too slowly and because Borrelia are extremely demanding to complexity of the growing substrate. Hence there is not any commercially available culture test for Lyme disease (Lyme disease org., 2012).

PCR (Polymerase Chain Reaction) multiplies a key portion of DNA from Borrelia that it can be detected. While PCR is highly accurate, it produces many false negatives because Borrelia are sparse and may not be present in the sample (Lyme disease org., 2012).

ELISA (Enzyme-Linked ImmunoSorbent Assay) and Western blot measure the patient's antibody response to infection. When a body is invaded by Borrelia, the immune system makes antibodies to fight the infection. ELISA is designed to be very "sensitive," meaning that almost everyone who has Lyme disease (and some people who do not) will get positive result. Furthermore, ELISA test is relatively cheap and easy to interpret. Hence it's the most commonly used test. On the other hand Western blot test is designed to be "specific," meaning that it is usually positive only if a person has been truly infected by Borrelia. However, Western blot is more expensive than ELISA (National Institute of Allergy and Infectious Diseases, 2012).

Because Lyme disease tests have low reliability doctors have to often rely on symptoms. Lyme disease has a myriad of symptoms. But the only specific symptom is erythema migrans, a circular, outwardly expanding rash. Unfortunately, only around 80% patients with Lyme disease develop the erythema migrans. And the 20% without the erythema migrans are often miss-classified (Aucott, 2012). Other Lyme disease symptoms include sinusitis, stiff neck, sweat attacks, muscle twitches, muscle weakness, involuntary jerking of limbs, arthritis, Bell's palsy, cramps, paralysis, depression, brain fog, insomnia, balance problems, light sensitivity, noise sensitivity, optic neuritis, nerve conduction defects, numbness, ECG (cardiac conduction) abnormalities, swallowing difficulties, tinnitus and more (Aucott, 2012).

1.3 Western Blot

An example of Western blots in Figure 1 displays two types of reactions– a reaction of Immunoglobulin G (in short IgG) to Borrelia (top) and a reaction of Immunoglobulin M (in short IgM) to Borrelia (bottom).

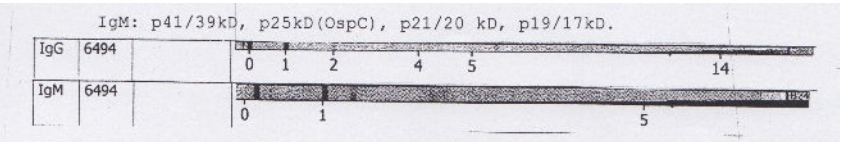

Figure 1: An example of Western blot test result for Lyme disease. Immunoglobulin G test is at the top; Immunoglobulin M test is at the bottom.

IgM antibodies are the first antibodies to be produced in the body in response to an infection and are produced in great quantity. IgM antibodies are large, up to six times larger than the IgG antibodies. IgM antibodies, when present in high numbers, represent a new active infection or an existing infection that has become reactivated. Over time, the number of IgM antibodies declines as the active infection is resolved.

IgG antibodies are produced once an infection has been going on for a while, and may be present after the infection has been resolved. Generally speaking, the presence of IgG antibodies to an organism when accompanied by a negative IgM test for the same organism means that the person was exposed to that organism at one time and developed antibodies to it, but does not have a current active infection of that organism. However, when it comes to Borrelia that is not necessarily the case, as Borrelia is excellent in hiding. Borrelia can repeatedly hide from immune system by altering its surface proteins, turning itself into undetectable cyst or moving into the brain, where it takes time to discover the infection. And then Borrelia can hit again (Kaplan, 2013).

To recapitulate:

- IgM is a sign of a current infection.
- IgG is a sign of a current infection, or of a past exposure to or past infection by the organism.

From the image processing view a Western blot (also called strip) is a white paper with dark vertical bands. The dark bands labeled as 0 and 1 are calibration bands.

The band 0 marks the beginning of the strip and the distance between the band 0 and the band 1 tells us whether it's IgG or IgM strip. Other labels mark positions of antigens typical for Lyme disease. For example the band 2 in IgG blot in Figure 1 could suggest that the patient is Lyme positive. There is also a band after the 5th label. However, this band is not specific for Lyme disease and as a consequence misses the label. The numbering of the labels is specific for the laboratory. However, universal labeling exists and uses molecular weight of the antigens in kilo-Daltons (kDa).

But not only that the position of the bands is evaluated but also their intensity. The band has to be as dark as the control band to be considered valid. Hence the strips in Figure 1 would be said to be negative by the automated evaluation in a laboratory.

1.4 Western Blot Interpretation

The old standard criteria (used by most laboratories and health insurance companies) are as follows: An IgM blot is considered positive if two of the following three bands are present: 24 kDa (OspC), 39 kDa (BmpA) and 41 kDa (Fla). An IgG blot is considered positive if five of the following ten bands are present: 18, 21 (OspC), 28, 30, 39 (BmpA), 41 (Fla), 45, 58, 66 and 93 kDa (Bogen, 1994).

However, this standard interpretation schema is criticized because it omits other significantly specific bands like 31 kDa (OspA) and 34 kDa (OspB) that are present in both IgM and IgG (Pavia, 1998). The modified schema that includes these two additional bands is used for example by IGeneX laboratory. Still there are over thirty additional weakly specific bands that are excluded from the test interpretation.

It can be assumed that these mentioned interpretation criteria were selected to be easily understandable and executable by a human. But if we didn't require interpretability by a human, we could likely get higher prediction accuracy with computers. To name just two limitations that can be lifted by computers: The bands don't have to be equally weighted anymore but each band can have it's own weight. And the interpretation doesn't have to be linear but interactions between the bands can be considered. And indeed the interactions between the bands are proved to be significant (Pavia, 1998).

1.5 Western Blot Normalization

Reproducibility and quantification of Western Blots is extremely difficult as WBs are based on an enzymatic reaction that is highly dependent on timing and exposure. And even thought variables like granularity and amount of the sample substrate, development time, length of exposure to primary and secondary antibodies, length of exposure to chemoluminescence and washing are corrected, significant variability in WBs is still present. Normalizing WBs against an internal control called "cut-off strip" decreases the variability in WBs (see Figure 2). However, these cut-off strips are unavailable.

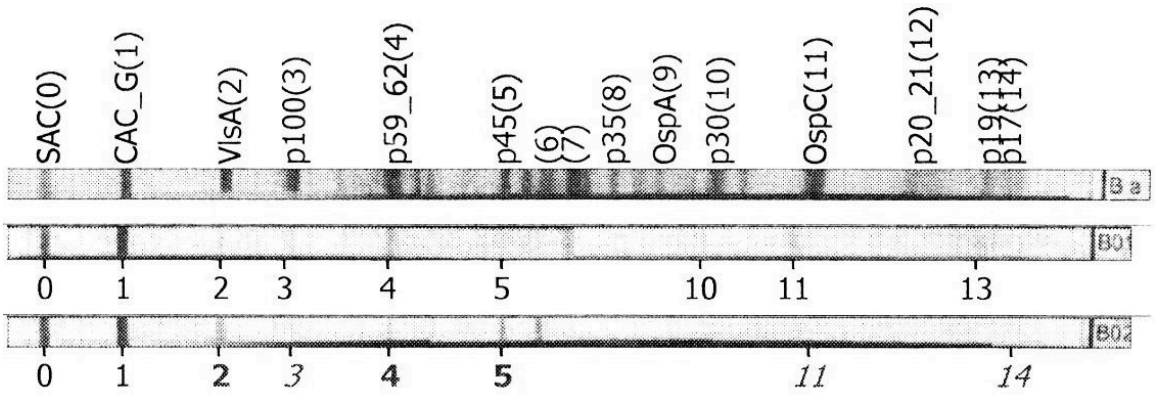

Figure 2: The cut-off strip (the strip at the top) works as a reference for the Western blots in a batch. The sample is provided by MUDr. Radek Klubal.

Absence of the cut-off strips means that we can't get the absolute amount of antibodies from the intensity of the bands anymore. In the case of evaluation of a single Western blot it would present a problem. However, we work with a set of Western blots and according to the law of large numbers, the average of the results obtained from a large number of Western blots should be close to the expected value, and will tend to become closer as more Western blots are examined.

Another complication is that the position of the bands is changing strip from a strip. Hence it is difficult for a machine to tell which band is which (see Figure 3 for demonstration of the strip variability).

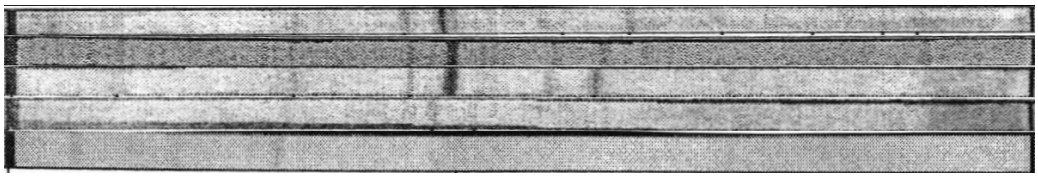

Figure 3: Intensity and position of the bands is changing batch from batch. The picture shows the Western blots from a single patient in process of patient's treatment.

Nevertheless the problem of moving peaks in a signal is also present in other fields like speech recognition or chromatography. And the problem was solved with Dynamic Time Warping.

Dynamic time warping (DTW) is an algorithm for measuring similarity between two sequences, which may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video the person was walking slowly and if in another he or she were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. An example of two signals aligned with DTW is in Figure 4.

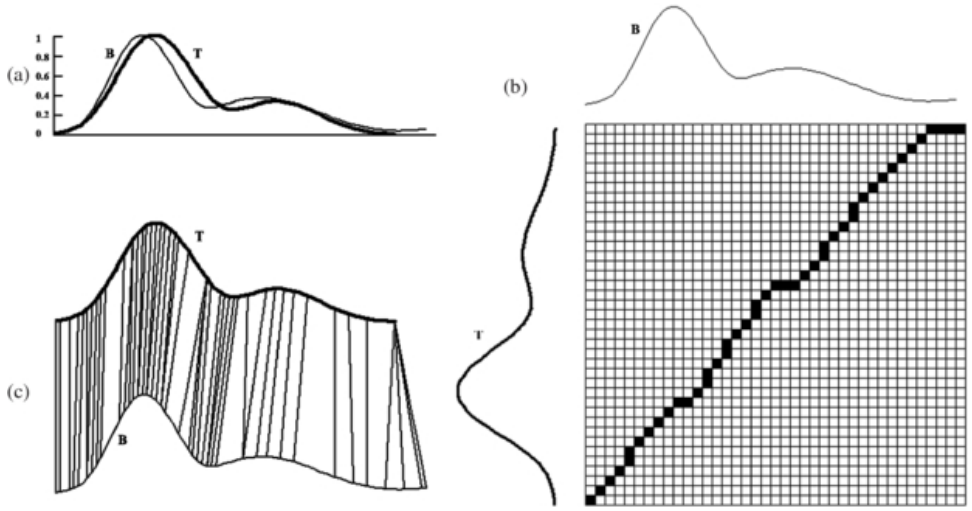

Figure 4: (a) While the two signals B and T have an overall similar shape, they are not aligned. (b) Dynamic Time Warping finds the ideal warping path. If the signals were identical, the path would be straight diagonal line. (c) The result of warped signals (Keogh, 2013).

However, if Dynamic Time Warping is left unconstrained it can stretch and contract the signals too much. To limit DTW elasticity two approaches are commonly applied. Either we limit the maximal shift of the two signals at any point to a constant value, that's called Sakoe-Chuba Band, or we can let the bound gradually increase from the ends, that's called Itakura Parallelogram (see Figure 5).

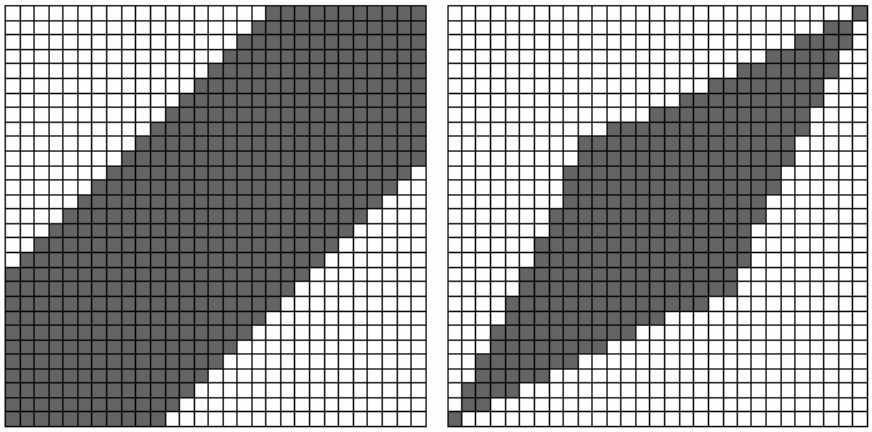

Figure 5: Two constraints: Sakoe-Chuba Band (left) and Itakura Parallelogram (right), both have a width of 5 pixels (Keogh, 2013).

1.6 Western Blot Registration

Because the Western blots are manually glued to the laboratory reports, their positions vary. There are four groups of algorithms for registration of rectangle objects: based on corner locations, based on edges, morphological transformation and Haar's detectors. The first three approaches share the same common characteristic – the rectangles are not identified directly. Instead a simple intermediate step is per-

formed, which drastically limits the amount of locations, where the rectangle can be found. The intermediate step can be for example corner detection. Once corners are found the algorithm walks thru the corners and attempts to find four corners that have the right orientation and position that they define a rectangle (de la Escalera, 1997). Another intermediate step can be edge detection. Once the edges are found the algorithm walks thru the edges and finds edges defining a rectangle (Jung, 2004). And morphological transformation simplifies the image into black-and-white image, where rectangles have one color and everything else have the second color (Faradji, 2007).

But Haar's detector works differently. It places a simple mask over the image and calculates the match. Based on the level of the match, the algorithm either tries different position with a better prospect, calls a match, or calls a failure to find the desired object. And while Haar's detector was originally developed for face registration (Viola & Jones, 2001) but it can also be used for general rectangle registration (Choi, 2007).

Another way how to find a rectangle object is to use template matching. In template matching we have a template image and finds the best match in the picture. The state of the art algorithm for template matching is Fourier-Mellin transformation, which can deal with rotation, shift and scale deformations. Other algorithms find characteristics features, like corners (Harris algorithm, "FAST" - Features from Accelerated Segment Test, the minimum eigenvalue algorithm developed by Shi and Tomasi) or other stable objects (Maximally Stable Extremal Regions, "SURF" - Speeded-Up Robust Features). Then the algorithm runs RANSAC algorithm, which robustly matches the points of interest in the template with the image. Because these algorithms are highly sophisticated, it was considered an overkill to use them for rectangle detection.

Out of the set trivial rectangle registration algorithms (based on corner, edge or morphological transformation) edge based rectangle registration is likely the most robust (Basile, 2010). Unfortunately, based on the performed experiments the vertical sides of Western blots are too short for reliable edge detection and consequently these algorithms fail. Hence corner detection and morphological transformation are left. Because of previous experience with cell counting in microscope photography via morphological transformation (Motl, Cell counting, 2011), which outperformed a commercial application (GSA, 2010) specializing in cell counting, morphological transformation method was used.

1.7 Machine Learning

Machine learning deals with construction and study of systems that can learn from data. In our case, a machine learning system could be trained on Western blots to learn to distinguish between Lyme positive and negative. After learning, it can then be used to classify new Western blots into Lyme positive and negative groups.

Based on the type of available data, there are two categories of machine learning algorithms: *supervised* and *unsupervised*. Supervised learning generates a function that maps inputs to desired outputs (also called labels, because they are often pro-

vided by human experts labeling the training examples). In our example, the supervised algorithm (called classifier) would be provided with labeled Western blots, where label would tell us whether the patient is Lyme positive or negative. And based on these training data the classifier would learn how to distinguish new Lyme positive/negative patients just by looking to their Westerns blots.

In the case of unsupervised learning, labels are not known during training. In our example, the algorithm wouldn't know which Western blots belongs to Lyme positive patients and which to Lyme negative patients. But still, the algorithm could find well-defined clusters representing either a group of Lyme positive patients or a group of Lyme negative patients. We wouldn't just know which group each cluster represent.

Finally, semi-supervised learning makes use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location). The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value.

1.8 Clustering

There are many different clustering methods, but three commonly used are k-means, EM-clustering and hierarchical clustering. Because hierarchical clustering provides the widest range of visualizations (dendrogram, clustergram, heat map) and is frequently used by bioinformatics, it was decided to use this method primarily.

1.8.1 Hierarchical clustering

Given a set of N items to be clustered, and a N×N distance matrix, the basic process of a basic hierarchical clustering can be described in following steps:

1. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances between the clusters equal the distances between the items they contain.
2. Find the closest pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3. Compute distances between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Step 3 can be done in different ways, which is what distinguishes single linkage from complete linkage and average linkage clustering.

1.8.1.1 Single linkage

In single linkage clustering (also called the nearest neighbor), we consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster:

$$d(r,s) = min(dist(x_{ri}, x_{sj})), i \in (i,..,n_r), j \in (1,..,n_s)$$

This linkage method produces clusters in shape of a chain.

1.8.1.2 Complete linkage

In complete linkage clustering (also called the furthest neighbor), we consider the distance between one cluster and another cluster to be equal to the longest distance from any member of one cluster to any member of the other cluster:

$$d(r,s) = max(dist(x_{ri}, x_{sj})), i \in (i,..,n_r), j \in (1,..,n_s)$$

This method performs well when clusters are rounded. But this method fails whenever clusters are elongated. Furthermore, this method has difficulties to create clusters of different sizes whenever clusters are too close together, because the smaller cluster steels members from the bigger cluster.

1.8.1.3 Average linkage

In average linkage clustering (also called UPGMA from *Unweighted Pair Group Method with Arithmetic mean*), we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster:

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj})$$

This method works equally well on both, rounded and elongated clusters. However, this method is more computationally demanding than previous methods.

We can also change distance metric. Distance metric expresses similarity of the two data points. Depending on what aspect of the data do we want to capture within the clusters, we can define different distance measures:

1.8.1.4 Euclidean distance

The Euclidean distance between points p and q is the length of the line segment $\overline{pq}$. If $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$ are two points in Euclidean n-space, then the distance from p to q is given by:

$$d_E(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

This metric describes "ordinary" distance between two points that one would measure with a ruler. To speed up calculation unsquared form can be used since it results into the same clustering.

This metric has certain advantages (e.g., the distance between any two objects is unaffected by other objects). However, Euclidean metric is not scale independent. For instance, if one dimension was in meters and another in millimetres, the dimension in millimetres would be weighted 1000 more then the dimension in metres. Hence it is generally good practice to normalize data to the same variance before using Euclidean metric (StatSoft, Inc., 2009).

1.8.1.5 Manhattan distance

The Manhattan distance (also known as city-block distance) between two items is the sum of the differences of their corresponding components:

$$d_M(p,q) = \sum_{i=1}^{n} |p_i - q_i|$$

This metric can be explained as the shortest path a car would drive in a city laid out in square blocks like Manhattan (neglecting that there are one-way streets and some oblique streets in Manhattan).

In most cases Euclidean and Manhattan metrics yield similar results. However, the effect of a single large dimension in Manhattan metric is dampened because the values are not squared.

1.8.1.6 Correlation distance

The correlation distance of two variables is obtained by dividing their distance covariance by the product of their distance standard deviations:

$$d_C(p,q) = \frac{cov(p,q)}{\sqrt{var(p) \cdot var(q)}}$$

This metric measures a similarity in shape of two expression patterns but is insensitive to the magnitude of the change. For instance, if we had a sine wave with amplitude of 1 and another sine wave of identical phase but amplitude of 10, they still would be considered to be identical. This insensitive to the magnitude of the change is useful for comparing signals, which weren't recorded in absolute scale.

1.8.2 Self Organizing Map

A self-organizing map (SOM) is a type of neural network that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map. Self-organizing maps are different from other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space. This makes SOMs useful for visualizing low-dimensional views of high-dimensional data. A self-organizing map consists of components called nodes or neurons. Associated with each node is a weight vector of the same dimension as the

input data vectors and a position in the map space. The usual arrangement of nodes is a two-dimensional regular spacing in a hexagonal or rectangular grid (see Figure 6 for illustration). The self-organizing map describes a mapping from a higher dimensional input space to a lower dimensional map space. The procedure for placing a vector from data space onto the map is to find the node with the closest (smallest distance metric) weight vector to the data space vector.

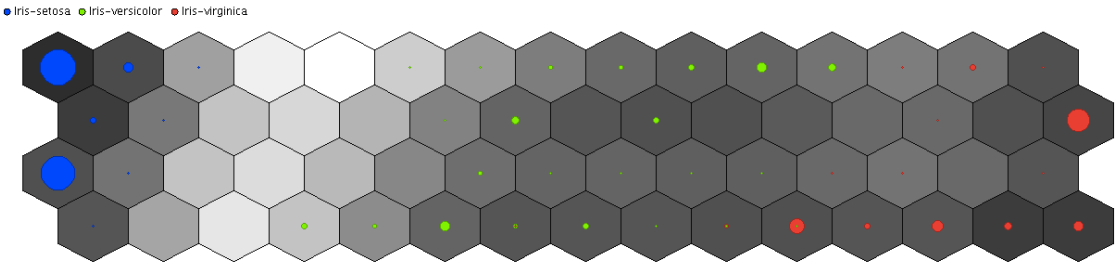

Figure 6: Iris dataset clustered with Self Organizing Map in own SOM plugin for RapidMiner.

2 Data Mining

Data mining project is a process, which requires a lot of resources. Beginning with human, over material to software. The common identifier of these resources is that they consume money. One approach how to save these resources is to perform the project in a standardized way.

For this purpose companies like SPSS and Teradata designed data mining process called Cross Industry Standard Process for Data Mining, in short CRISP-DM. The data mining process is depicted in Figure 7 and is de facto standard for developing data mining and knowledge discovery projects among data miners (Óscar Marbán, 2009).

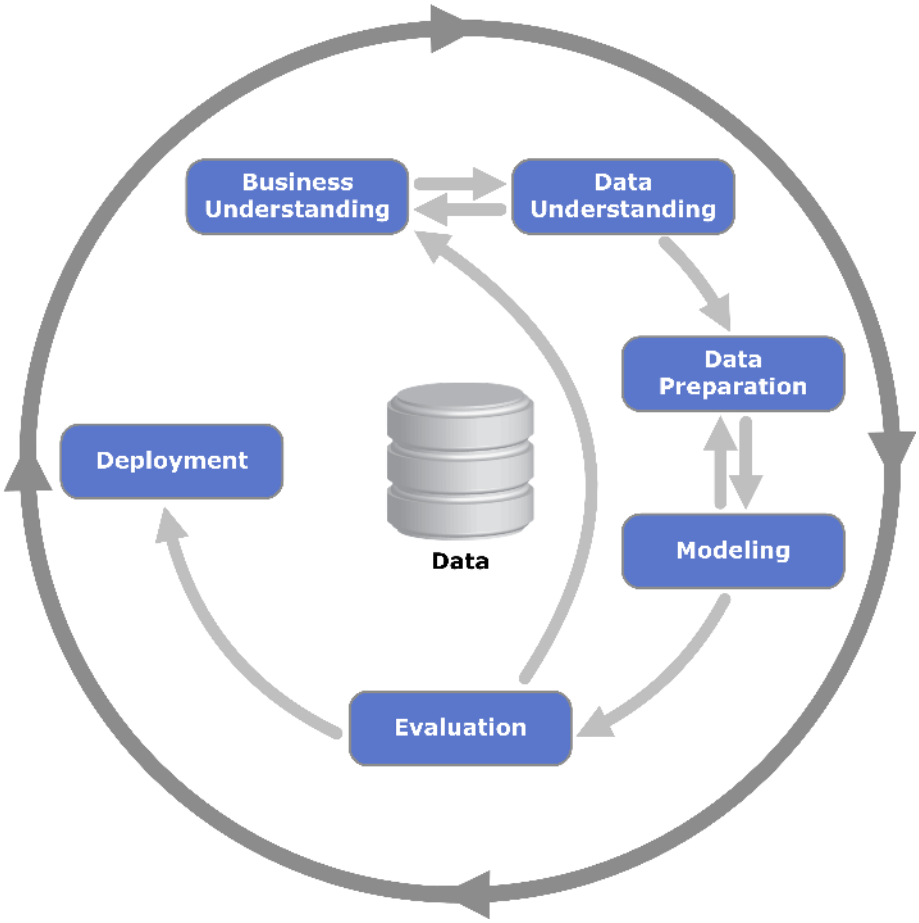

Figure 7: Process diagram showing the relationship between the different phases of CRISP-DM (Wikimedia Foundation, Inc., 2013).

However, CRISP-DM diagram doesn't illustrate time consumption of each phase. Hence a modified schema depicting time investment into a data mining project from a data mining consultant Dr. Zdeněk Skála is used (Figure 8) to describe the thesis structure.

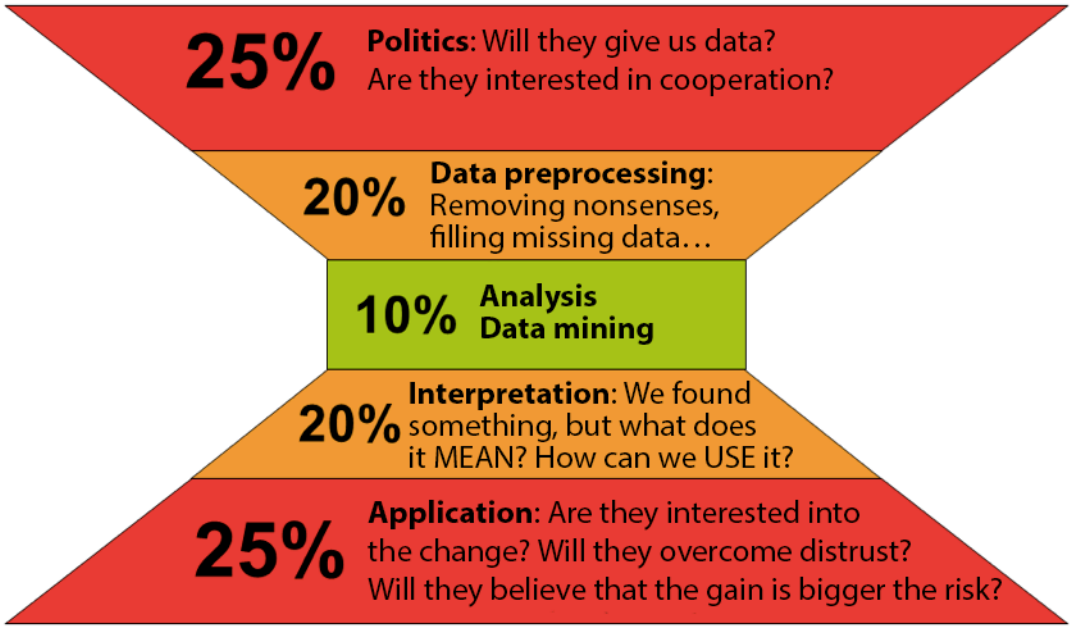

Figure 8: Time investment into a data mining project.

The project begins with obtaining data. It took two months to convince the doctor that the project is viable and meaningful. And then it took another month to manually scan over 4227 pages. This phase is described in *data acquisition* section.

In the following step, described in *data preprocessing* section, Western blots had to be located on the scanned pages and converted into the form usable by machine learning methods. This step also took a lot of time since process had to be automatized because of the sheer volume of data (next to 4227 pages another 22000 pages are waiting in the archive and new pages are coming each day). Furthermore a new preprocessing workflow had to be developed since in all examined publications (the list of the related publications is in appendix A) the authors were using raw Western blots while this study uses copies of Western blots.

The last step performed in the thesis is analysis of the preprocessed data and is described in *knowledge discovery* section.

2.1 Data Acquisition

In total 4227 Western blot reports were scanned. These reports were kindly provided by MUDr. Radek Klubal. The reports itself were produced by *Oddělení parazitologie, mykologie a mykobakteriologie Praha, Zdravotní ústav se sídlem v Ústí nad Labem.* Following data from the reports were collected: IgG and IgM Western blots with laboratory conclusion, ELISA test with the conclusion, date of accepting the blood sample and salted MD5 hash of the patients identification number (*rodné číslo*).

2.1.1 Western Blots

The Western blots were scanned with a high speed scanner Canon DR-M160 at 300 dpi with 24 bit color depth and stored into jpeg files with loss compression set to high quality (85%). Higher quality scans were also considered; however, performance of a system is generally limited by its weakest point. And in our case the limiting factor is the quality of the raw data - Western blots. The blots were scanned in the laboratory and printed out with a laser printer at 150 dpi as depicted in Figure 9. Hence by Nyquist–Shannon sampling theorem scanning with double resolution preserves all the details present in the blot.

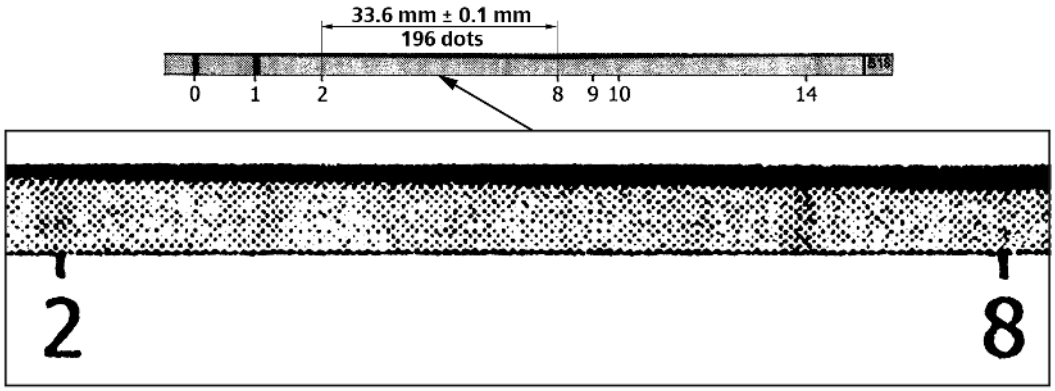

Figure 9: The Western blots are printed at 196/33.6*25.4 ≈ 150 dpi.

To examine the impact of lossy jpeg compression the same Western blot was scanned twice, once with lossless compression and once with lossy jpeg compression. Then the blots were averaged by columns to extract one-dimensional signals just like in the preprocessing step described further. The maximal difference in the signals is up to one degree of intensity as depicted in Figure 10. Since color images scanned at 24 bit depth are capable to display 256 shades of gray the difference represents a change of 1/256 ≈ 0.4% and as such is considered insignificant. The decision to use lossy compression reduced the storage requirements approximately five times.

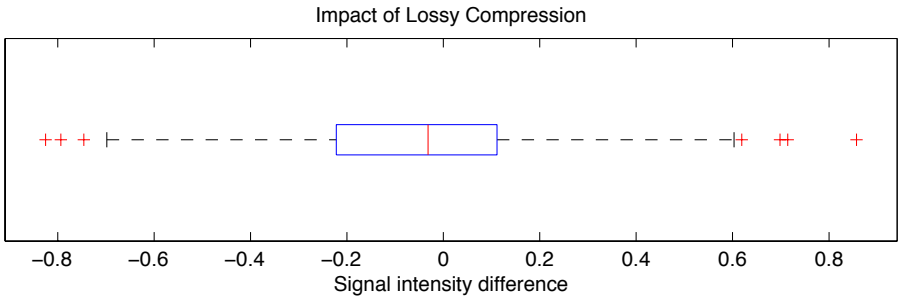

Figure 10: Impact of lossy jpeg compression to the average Western blot intensity value. The mean difference is 0.0456 and the maximal difference is 0.8571. The source image is the same as in Figure 9.

The laboratory conclusion whether the patient is positive or negative based on IgG and IgM blots were also collected for comparison of the proposed method with the current standard.

2.1.2 ELISA Test

ELISA tests were also collected because the available data didn't contain information whether the patient is truly positive or negative. Only in a limited amount of samples the doctor's best prediction was available. Hence ELISA test provides a verifiable hypothesis that the western blot prediction works – the WB positive group should have proportionally more ELISA positives than the WB negative group.

2.1.3 Time Identification

It's important to have a date signature of a test because we commonly have several tests from each patient. And the time signature allows us to order the tests in time and observe WB's reaction to Lyme treatment. There were three time signatures available – date of blood collection, date of accepting the blood sample by the laboratory and finally date of finishing the report. Theoretically any of these dates alone should be enough for unique test ordering. However, sometimes it happens that the report arrives in batches as each particular test can be actually performed at different places in different times. Therefore if we used date(s) of finishing the report(s), some reports would be artificially scattered in time and induce noise. Similarly if we used date of accepting the blood sample we would have a problem that each test can be executed by different laboratory and therefore the tests could have different pick up dates. While in our case both WB and ELISA were executed by the single laboratory, other potentially useful blood tests were performed in different laboratories. And if we wanted include results of these tests in our analysis in the future, we would again get meaninglessly scattered dates. Therefore we are left with date of blood collection. Since we may assume that a patient would object against frequent blood sampling, blood collection date has likely the lowest noise and is used in this analysis.

2.1.4 Patient Identification

Additionally some kind of patient's identification (ID) was required to identify tests belonging to one client. However, this ID can't be anything that would relief the connection between the collected data and real person because medical records are considered to be sensitive information and as such are protected by law. If sensitive information is to be made public, it has to be anonymised. Hence we have two conflicting requirements – we need unique patient's identification while the patient's identity must be obscured. A possible solution to this conflict is to use hash. Hash provides one way encryption. That means that it's easy to encode the protected message but it's hard, if not impossible, to decode the message back into the original state. Hence hash provides patient's obfuscation. Additionally, hash has to produce different outputs for different inputs but the same output for the same input passed

to the hash function twice. Therefore hash also provides unique patient's identification.

It could be argued that a plain table associating patient's identification with a serial number from 1 to the number of patients would be enough. However, if this table leaked the collected data would be relieved. The implemented solution doesn't contain any such table and is therefore immune to this vulnerability.

However, hashes bring a different type of vulnerability – brutal force and rainbow attack to the hash. The idea of the brutal force attack is to generate all possible *rodná čísla* one by one, hash each *rodné číslo* and compare the output with the hash to crack. Once the hashes equal, the attacker knows the *rodné číslo* representing the hash and moves to another hash. The rainbow attack accelerates hacking by storing precomputed pairs of *rodné číslo* – hash. And when an attacker wants to decode hash he/she just look it up in the precomputed table.

These types of attack are possible because the structure of *rodné číslo* is well known and there are only around 100 million combinations to try. To deal with this vulnerability additional text, called salt, was appended behind each *rodné číslo.* Because this salt is secret, the attacker would now have to generate all *rodná čísla* appended by all possible salts. This makes brutal force and rainbow attacks practically unfeasible.

The type of collected information and the anonymization process was discussed and approved by *Úřad pro ochranu osobních údajů*. The approval letter is in Appendix B.

2.2 Data Preprocessing

It would be an overkill to teach machine learning algorithms how to process whole Western blot reports. Instead Western blots are preprocessed to simplify machine learning process as much as possible. As illustrated in Figure 11 the first preprocessing step is registration of the Western blots from the scanned reports. While it could have been an easy task if the blots' locations were fixed on the report, the blots are glued to the report manually and their positions, rotations and even scales vary. Hence advance image processing tools are employed in WB registration. Once the WBs are extracted from the reports, they have to be normalized because bands shift uniquely in each WB. Once the WBs are normalized the WBs are converted into signals. This conversion greatly reduces storage and computation time because it converts 2 dimensional images into a smaller 1 dimensional signal. And still it preserves all useful information present in the WB.

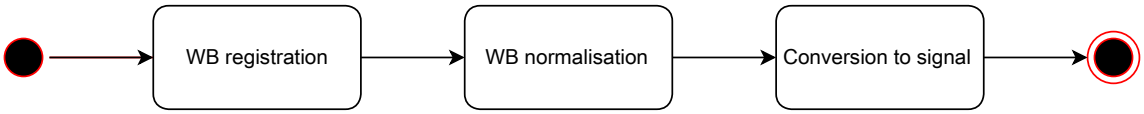

Figure 11: Activity diagram of Western blot (WB) preprocessing. WB report is processed into signal usable for machine learning algorithms.

Unfortunately, whole preprocessing had to be designed from scratch because all examined studies of Western Blots were using original WBs. As a consequence their preprocessing steps were quite different. To name a few principal differences the

size and position of WBs on the originals is fixed while in our data these things varies. Similarly noise is basically absent in the originals while we had to pay attention to this detail. However, the decision to use data from a doctor and not from a laboratory has some advantages. Namely, the doctor has access to much broader range of information about the patient than a laboratory. The laboratory knows just elementary patient's identification, his/her insurance company and the result of their own tests. But the doctor has access to the test results from all different laboratories, the patient's symptoms and the patient's anamnesis. Additionally, in this case the difference is also in the count of analyzed Western Blots as depicted in Figure 12.

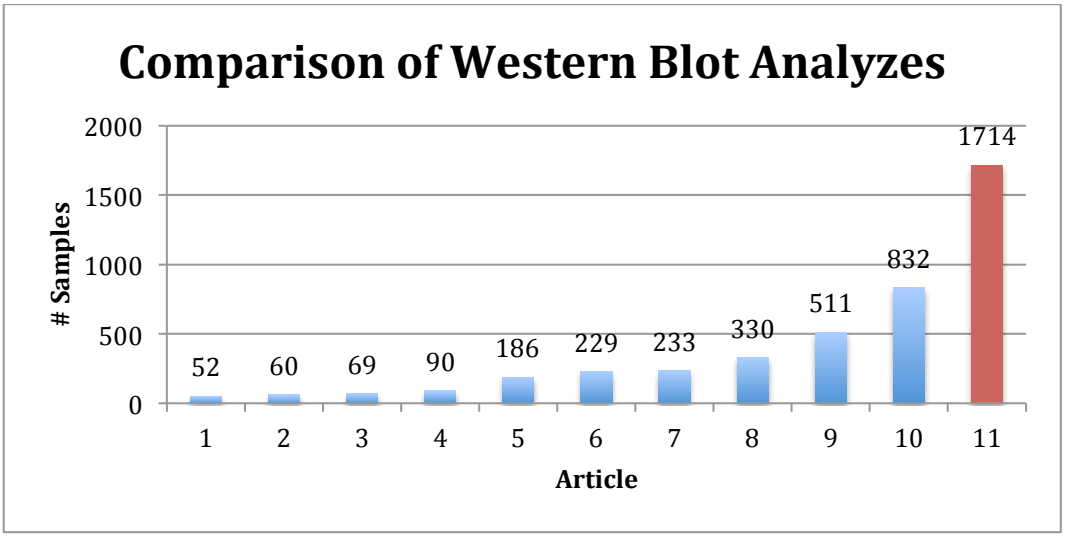

Figure 12: Comparison of articles analyzing Lyme disease Western Blots based on the count of Western Blots analyzed. The list of the examined articles is in the appendix. The red column depicts number of Western Blots analyzed in this thesis.

2.2.1 Western Blot Registration

Western blot registration is based on morphological rectangle detection and is depicted in Figure 13.

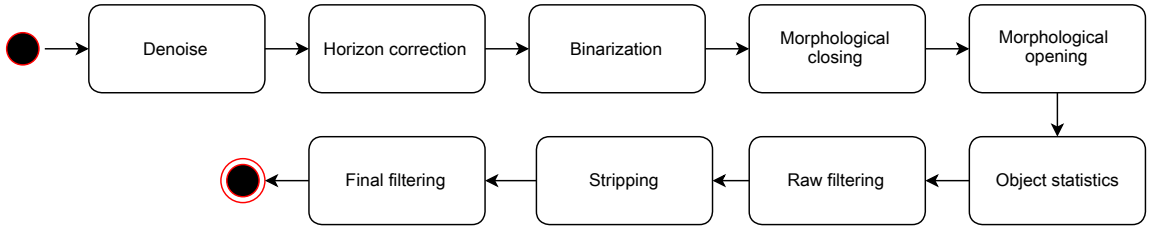

Figure 13: Activity diagram for Western blot registration.

The first step in image processing is commonly denoising because digital images are hardly ever completely noise-free and removing of the noise at the beginning of the processing chain can greatly simplify design of the following processing blocks (Buades, 2013).

The second step is automatic page alignment because the scanned pages can get slightly rotated during the scanning due to imperfect paper feeding. Despite exten-

sive literature check and consultation with image processing specialist (Ing. Daniel Sýkora, Ph.D. from Czech Technical University) no algorithm for automatic text rotation correction was found in the time of writing this thesis. Hence three different algorithms were designed, implemented and tested.

The next step is binarization of the image, which separates WBs from the background. Because the used sheetfed scanner produced equally illuminated digital images, simple global thresholding technique is sufficient for this task. Different thresholding algorithms were preliminary tested in ImageJ application on the sample reports. And Triangle thresholding seemed to perform the best in this task from seventeen different algorithms implemented in ImageJ (Landini, 2013). To have a comparison with the state of the art thresholding algorithm, performance of Triangle method was compared to Otsu's method. The binarization block can be theoretically placed at the end of the processing chain; however, the reduction of the 24-bit image to 2-bit image means great acceleration of the following processes. Hence binarization is placed just behind horizontal correction, which performs better on grayscale images than black and white images.

Morphological transformations like morphological opening and closing reduce the complexity of the objects in the image. Morphological closing removes small holes in the WBs while opening removes small objects like thin lines around the WBs (see Figure 1 for illustration of mentioned imperfections in WBs).

After the morphological transformation each separate object is identified and enclosed in a rectangular envelope. Then objects too small to be a WB are removed. This filtering greatly reduces total processing time of the following function, which precisely extracts strips from the enclosing envelope. After the precise extraction the objects are filtered once again. The detail description of each block follows.

2.2.1.1 Denoise

Whenever the source image is damaged by noise it's common to start the preprocessing step by denoising. In our case we work with images that were scanned in the laboratory, printed out and then again scanned in the doctor's ordination. Hence the digitized images contain quite a lot of imperfections. The small imperfections like wood vessel segments on the paper surface can be easily removed by denoising. Among the simplest and still effective denoise filters is median filter, which can suppress isolated noise without blurring sharp edges. Specifically, the median filter replaces a pixel by the median of all pixels in the neighborhood:

$$y[m,n] = median\{\, x[i,j], (i,j) \in\ w \,\},$$

where w represents a neighborhood centered around location (m,n) in the image.

Another popular noise reduction filter is Wiener filter. Weiner filter performs deconvolution of an image that was convolved with known (or estimated) kernel and also degraded by an unknown additive noise. Hence Wiener filter is used for deblurring of images and/or noise removal.

The third tested method is unsharp mask filter because this method was found superior to median filter and Wiener filter in a comparison of denoise filters applied to

medical images (Mahmood, 2011). Unsharp mask filter extracts edges from the image and then it sums the edge layer with the original image, making the bright edges brighter and dark edges darker but otherwise leaving the image intact. The name of unsharp mask comes from how the edge layer can be obtained - by blurring the original image and subtracting the blurred image from the original image.

2.2.1.2 Horizontal Correction

Three types of horizontal correction approaches were designed and tested: Fourier transformation, Hough transformation and dominant element orientation.

Fourier transformation technique is illustrated at a binary image of English text in Figure 14.

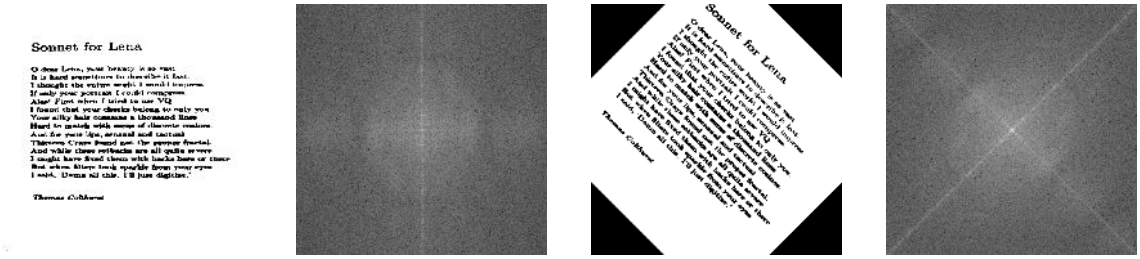

Figure 14: From left to right: the original text from Robert Fisher (Fisher, 2003). The logarithm of the magnitude of the original text's Fourier transform. The original text rotated by 45°. The logarithm of the magnitude of the rotated text's Fourier transform.

Two-dimensional Discrete Fourier transformation converts the image into frequency domain. Then the logarithm of the magnitude is maximal in the direction perpendicular to the dominant orientation of the text in the original image. If we proceed in the same way with the text, which was rotated about 45°, we can see that the line of the main peaks in the Fourier domain is rotated according to rotation of the input image. The second line in the logarithmic image (perpendicular to the main direction) originates from the black corners in the rotated image.

The magnitude image gives us information about the text's orientation in -90° to 90° range (the magnitude image is symmetrical hence it doesn't cover whole 360°). However, for rotation invariant rectangle detection it is sufficient to rotate the image in -45° to 45° range to guaranty that the blot's edges will be (approximately) vertical and horizontal. If we assume the majority of objects in the image have right angles, we average values at $mod(k, 90)$ increasing signal-to-noise ratio. This was implemented by folding the magnitude image along one of the axis passing thru the images center and averaging the two folds.

Once we have folded magnitude image we convert the image from Cartesian to polar coordinates and the resulting image is summed along the rows. The row with the highest sum is the orientation of the original image.

Hough transformation finds lines in the image as illustrated in Figure 15.

Figure 15: From left to right: original image from Robert Fisher (Fisher, 2003), edges from Canny edge detection applied to the original image, edges overlaid with lines identified by Hough transformation.

A naive approach to text orientation detection is to select the longest line. However, a more robust method is to create a weighted histogram of all lines' directions in the image. Luckily we don't have to compute the histogram because this information is already present in Hough transformation. Hough transformation describes the straight line position in polar coordinates: the horizontal axis defines the line's orientation and the vertical axis defines the line's distance from the origin. And the intensity of the point in Hough transformation defines the line's length. Hence all we have to do to find the dominant text's orientation is to find a column with the brightest points. Unfortunately we can't just average the intensity of all the points in the columns because this value is constant in the Hough transform. However, variance is a sufficient statistics to find the most dominant text orientation as depicted in Figure 16.

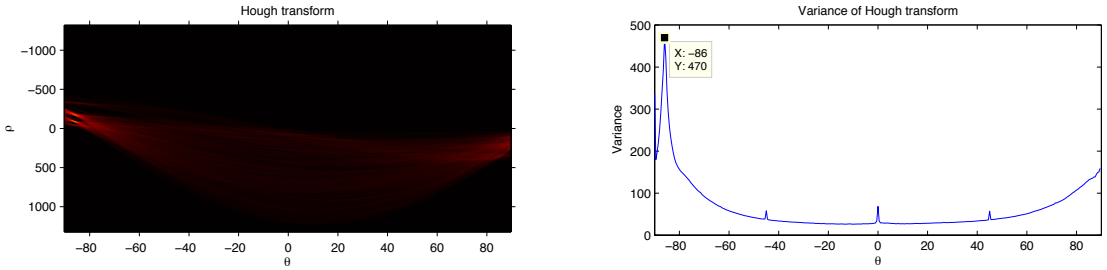

Figure 16: On the left is Hough transformation of the image in Figure 1 rotated by -86° from its vertical orientation. Right: the variance of the Hough transform peaks at -86°. That means the dominant orientation is 4° from the ideal horizontal orientation.

Some artifacts can be found in Hough transform at -45°, 0° and 45°. These artifacts are caused by discrete character of Hough transform. When continuous alternative to Hough transform, Radon Transform, is used these artifacts disappear. However, Radon transform is generally slower than Hough transform. Considering this tradeoff Hough transform is further used.

Furthermore the output of Hough transform can be also folded in to -45° to 45° range to further increase signal-to-noise ratio.

Dominant object orientation is obtained from binarized image with Otsu's method. Then the binary image is morphologically opened and closed to remove pepper and salt noise. Afterward each blob's orientation is calculated from the blob's pixel position:

$$uxx = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n}$$

$$uyy = \frac{\sum_{i=0}^{n}(y_i - \bar{y})^2}{n}$$

$$uxy = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$orientation = atan\left(\frac{uyy - uxx + \sqrt{(uyy - uxx)^2 + 4 * uxy^2}}{2 * uxy}\right)$$

where uxx, uyy and uxy are normalized second central moments for the blob. These orientations are counted in a histogram and the bin with the highest count identifies the text orientation. The method is illustrated on the example from Figure 1 in Figure 17.

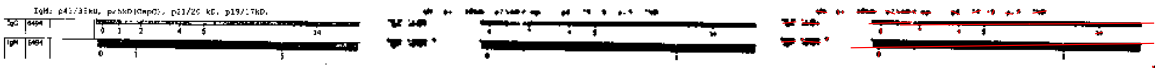

Figure 17: Process of image rotation detection with dominant object rotation. From left to right: binarized image, morphologically opened and closed image, image with blobs' orientation in red.

2.2.1.3 Binarization

Two types of thresholding algorithms were examined, traditional Otsu's method and Triangle method (Cytochem, 1977), which had promising preliminary results.

Otsu's method exhaustively searches for the threshold that minimizes the intra-class variance. On the other hand Triangle is based on geometry. Triangle method calculates histogram, finds the biggest peak in the histogram and draws a line from the peak to the opposite end of the histogram. Then it finds point A as shown in Figure 18 and adds a fixed offset.

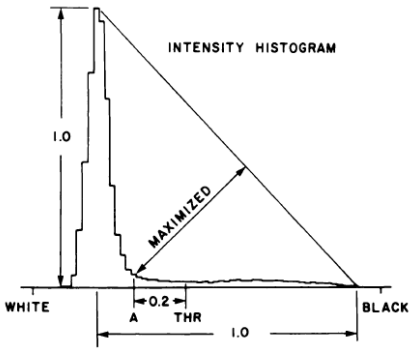

Figure 18: Triangle threshold method calculates histogram, finds the peak and draws a line from the peak to the opposite end of the histogram. Then it finds point A as shown and adds a fixed offset. The image is from (Cytochem, 1977).

2.2.1.4 Morphological Transformations

The scanned images contain many small and some bigger imperfections. The small imperfections like wood vessel segments are filtered out by denoise filters. But bigger imperfections like human hair, pencil marks and paper tears (see Figure 19 for illustration) are better to remove with morphological transformations.

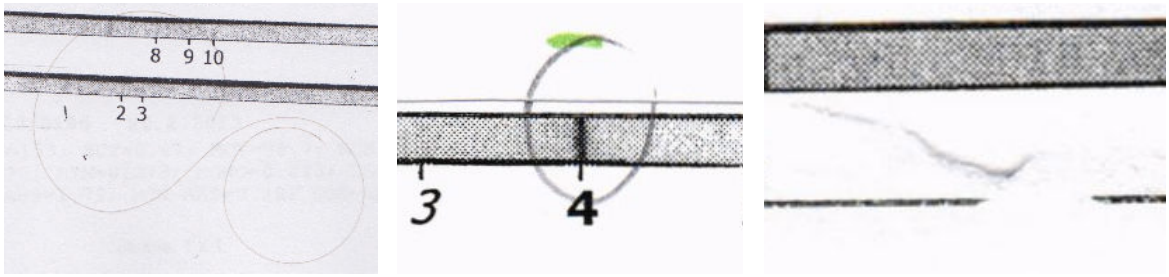

Figure 19: The scanned images contain imperfections like human hair (left), pencil marks (center) and paper tears (right).

The underlying principle of morphological opening and closing is the same. Both morphological opening and closing are using two fundamental operations: erosion and dilatation.

The erosion operator takes two pieces of data as inputs. The first input is the image, which is to be eroded; the second is the kernel size. To compute the erosion of a binary input image by the kernel, we consider each of the *foreground* pixels in the input image in turn. For each foreground pixel (which we will call the *input pixel*) we superimpose the kernel on top of the input image so that the origin of the kernel coincides with the input pixel coordinates. If for *every* pixel in the kernel, the corresponding pixel in the image underneath is a foreground pixel, then the input pixel is left as it is. If any of the corresponding pixels in the image are background, however, the input pixel is also set to background value. Figure 20 illustrates the effect of erosion on a binary image with kernel size 3×3.

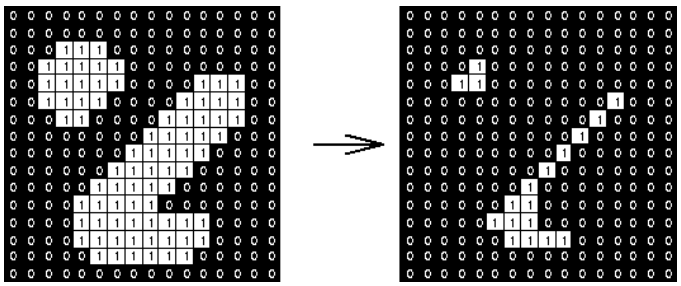

Figure 20: Effect of erosion using a 3×3 square structuring element (Fisher, 2003).

Dilatation works similarly. The kernel is slid over the image and if at least one pixel in the kernel coincides with a foreground pixel in the image underneath, then the input pixel is set to the foreground value. If all the corresponding pixels in the image are background, however, the input pixel is left at the background value. The effect of dilatation is depicted in Figure 21.

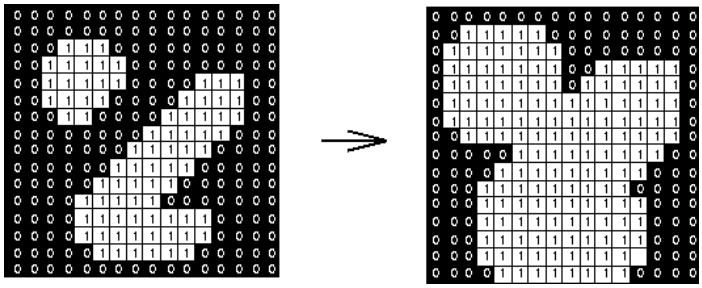

Figure 21: Effect of dilation using a 3×3 square structuring element (Fisher, 2003).

2.2.1.5 Filtering

We know following information about Western blots:

- typical width and height,
- typical aspect ratio,
- typical location on a page,
- blots are approximately aligned with the text.

To quantify this knowledge 65 randomly selected blots were manually measured. There are three typical Western blot widths as depicted in Figure 22. When this difference was discussed with the laboratory, it turned out that they were using different magnification setting during the scanning of the original Western blots. But I was assured by the technician in the laboratory that the Western blots and the process behind making them remained essentially the same.

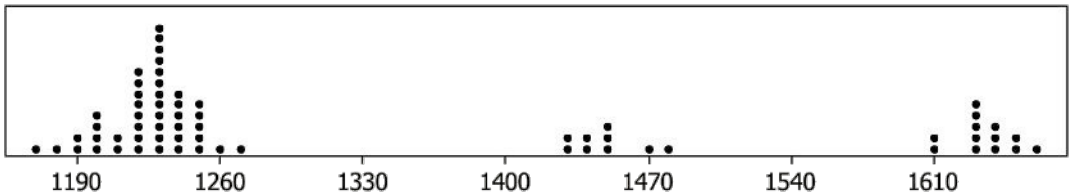

Figure 22: Dotplot of Western blot widths in pixels at 300 dpi. There are three characteristic sizes of Western blot widths.

Similarly Western blot heights vary greatly as depicted in Figure 23.

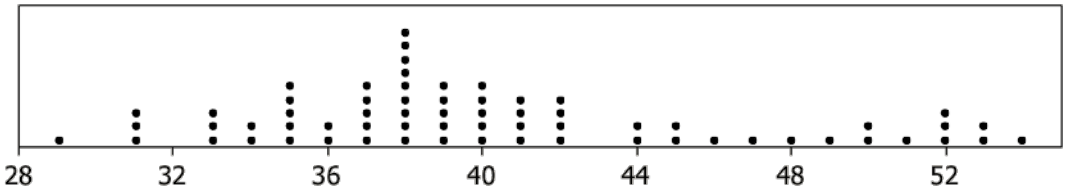

Figure 23: Dotplot of Western blot heights in pixels at 300 dpi.

The Western blots' area (height × width) is in a narrower range than if extreme heights and widths were multiplied as depicted in Figure 24.

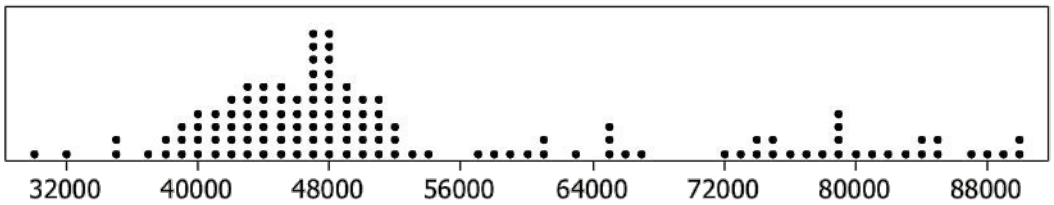

Figure 24: Dotplot of Western blot area in pixels at 300 dpi.

The aspect ratio of the blots is 33±9.

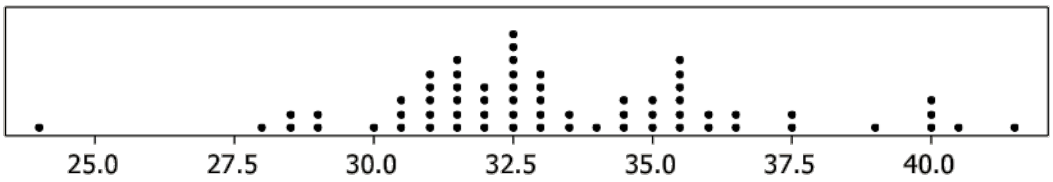

Figure 25: Dotplot of Western blot aspect ratios. The outlier at 24 is a valid measure.

The reports are printed on A4 pages, hence their size is 2 437 pixels to 3 524 pixels when scanned at 300 dpi. Based on the measured blots' positions on the reports depicted in Figure 26 and Figure 27 the top left corner of the blots is always in the bottom left half of a report.

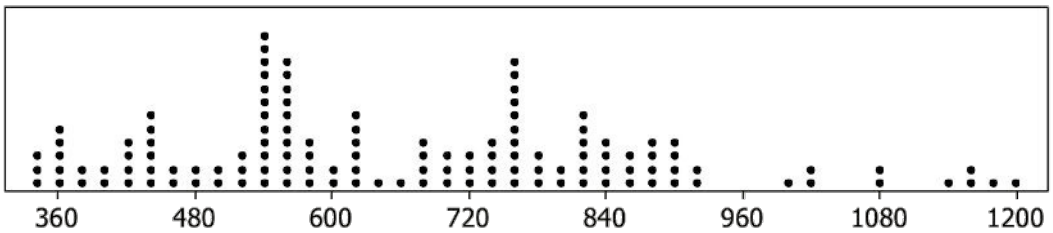

Figure 26: Dotplot of Western blot's top left corner positions on the report. The distance is measured in pixels from the left report's edge at 300 dpi.

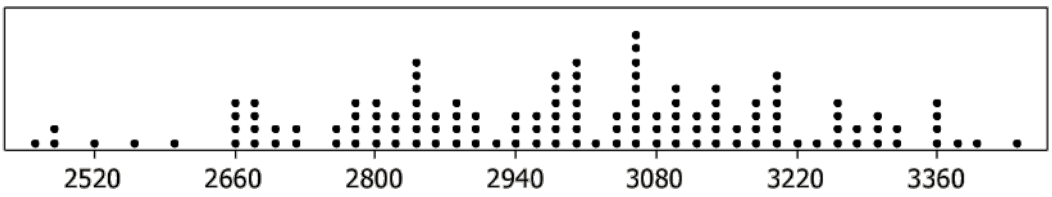

Figure 27: Dotplot of Western blot's top left corner positions on the report. The distance is measured in pixels from the top report's edge at 300 dpi.

The orientation of Western blots is 0 ± 5° as depicted in Figure 28.

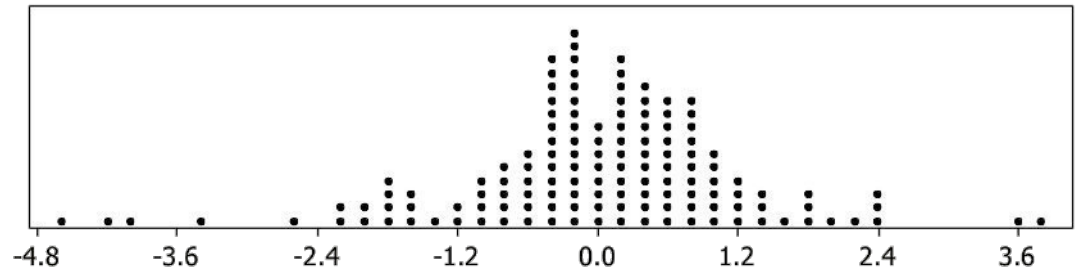

Figure 28: Dotplot of Western blot orientations in degree in respect to the page orientation.

Based on the collection of 65 real WBs and 35 objects falsely classified as WBs a decision tree depicted in Figure 29 was expertly constructed. Note that this filter tests WBs only on the lower size bound because the WBs have to be still shrink in the stripping phase. Similarly aspect ratio couldn't be applied here because of the current aspect ratios can drastically change. However, there is one decision block that wasn't so far discussed. Sometimes it can happen that a homogenous gray object on a scan is mistakenly classified as the dark object by the global image binarization. This issue can be either eliminated by using local image binarization or by checking difference between the maximal and the minimal pixel intensity and checking whether the difference is above the threshold. The last approach has an advantage that the calculation is typically performed only on a small part of the report.

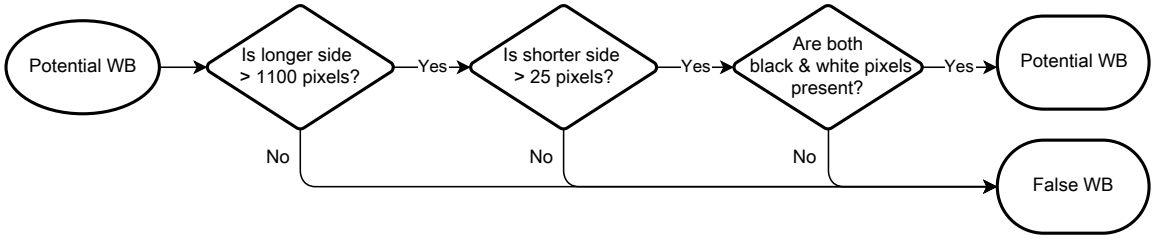

Figure 29: Decision tree for raw Western blot filtering.

2.2.1.6 Stripping

The blots are rarely exactly horizontal on the page because the blots are manually glued to the report. The stripping function takes the blot's bounding box (depicted in red in Figure 30) and applies the horizontal correction function to correct the declination.

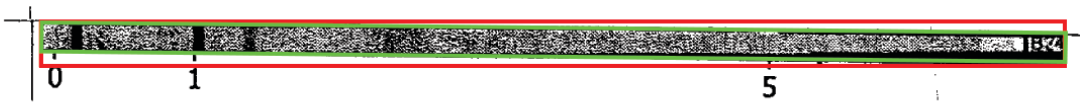

Figure 30: The stripping function takes rectangle bounding box of the strip (red) straightens the strip and then cuts away white parts (green).

Once the blot is horizontal the white space around the blot can be hungrily removed row by row or column by column till the mean intensity in the row/column is above the threshold as described by the following pseudocode:

```
1)  Blot = horizonCorrection(Blot)
2)  For Line = {Blot.top, Blot.bottom, Blot.left, Blot.right}
3)     Do
4)        If the mean value of the Line is above the threshold
5)           Remove the Line
6)           Line = next Line
7)        Else
8)           break
9)        End
10)    End
11) End
```

The output of the stripping function is depicted in green in Figure 30.

As discussed in the article about using morphological features for fast logo detection (Hassanzadeh, 2011) the biggest problem of the morphological detection are overlapping objects. In the example in Figure 31 two strips are placed over a stamp with signature. Because the stamp and the signature have noisy background, the threshold function interprets this area as a strip. Together with exaggerated morphological transformation the two strips blend together and strip registration consequently fails.

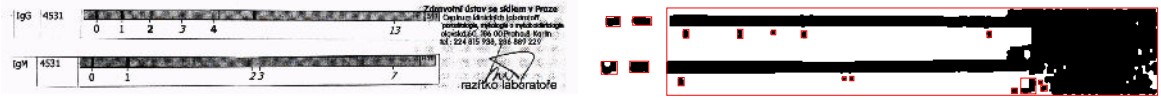

Figure 31: Example of the strip overlapping with the stamp (left). After binarization and morphological closing and opening with size 5×5 the two strips can blend into one object represented by the red bounding box (right).

To avoid the blending of the strips morphological transformations must be subtle. However, Figure 32 illustrates a case where strong morphological transformation is helpful because the printer "made white strips" on the printout.

Figure 32: Example of strips with vertical white strip caused by the printer (left). Morphological closing and opening with size 7×7 correctly reconnected the strips (right).

The problem is that overlaying objects require morphological transformations with size *smaller* 5×5, while white strip correction requires morphological transfor-

mations with size *bigger* 6×6. Hence we have conflicting requirements that weren't solved in Hassanzadeh's work.

Nevertheless the contradicting requirements on the ammount of morphological transformation can be solved by favouring one extreme and dealing with the broken case. The implementation in the thesis exagerates morphological transformation. This approach results in blended strips like in Figure 31. The blended strips are then recursively halved in horizontal direction and stripped until the height of each strip is below the threshold:

```
1) function Strips = halve(Strip)
2)    if Strip.height > Threshold
3)       upperHalf = stripper(Strip.upperHalf)
4)       lowerHalf = stripper(Strip.lowerHalf)
5)       Strips = [upperHalf, lowerHalf]
6)    else
7)       Strips = Strip
8)    end
9) end
```

2.2.1.7 Final filtering

The final filtering is similar to raw filtering. However, the strip's size and position are now final. Hence the strip's size can be upper bounded to 1 800 × 60 pixels. The strip's aspect ratio is checked to be in 20 to 50 range and the strip's area is checked to be between 25 000 to 95 000 pixels. The strip is checked to have the top left corner in the bottom left half of the report. Similarly a rotation filter was added, as strips are ± 6 degree horizontally placed. The diagram of the filter is in Figure 33.

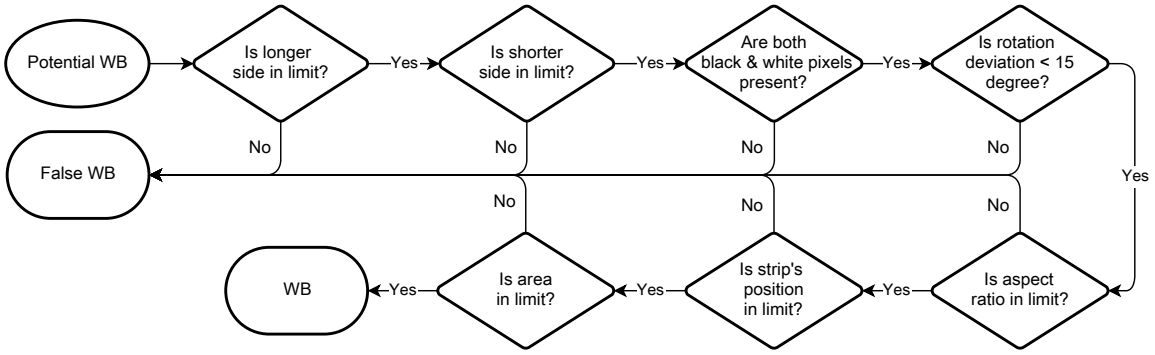

Figure 33: Decision tree for final Western blot filtering.

2.2.2 Normalization

As depicted in Figure 34 the sizes of the strips differ even though they were scanned in the same resolution. And common machine learning algorithms can't work with variable image length (with variable attribute number in machine learning terminology).

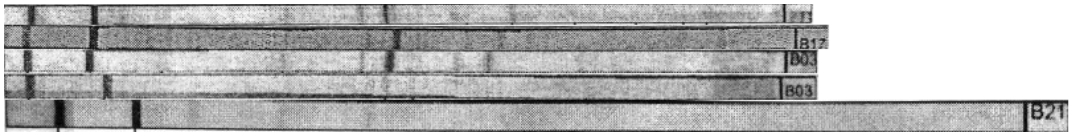

Figure 34: Even though all reports were scanned at the same resolution their sizes differ. The picture shows five IgG strips of one patient collected during the patient's treatment.

The solution to this problem is to rescale the strips to the average size. An illustration of the resized strips is in Figure 35.

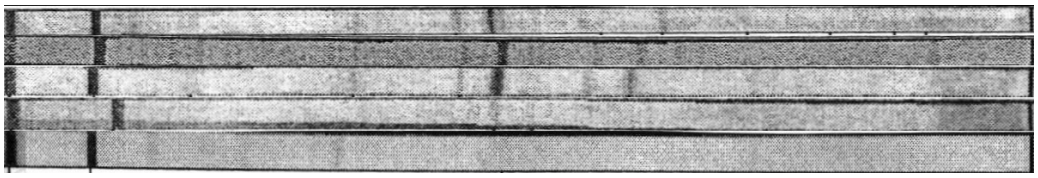

Figure 35: Each strip from Figure 34 was automatically shortened to show only the area within the first band and the right border. Then the strips were resized to the uniform size with bilinear interpolation.

The algorithm finds the first strong band from the left. Then it cuts away everything on the left from the first band. The similar procedure repeats on the right end. In the end the strips are resized to the same length with bilinear interpolation.

Unfortunately, even after this normalization the bands are still unaligned. Neither aligning to the second control band helps much as depicted in Figure 36.

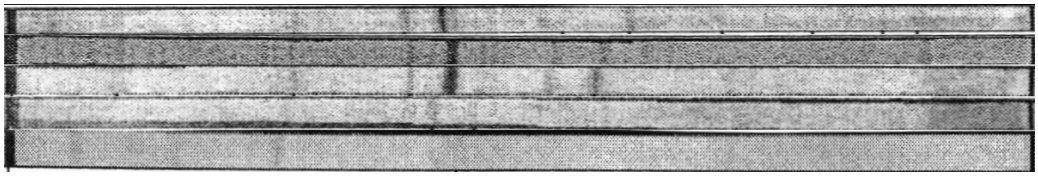

Figure 36: Each strip from Figure 34 was again automatically shortened to show only the area with bands but this time the left side limit is the second control band. Then the strips were resized to the uniform size.

Hence strips were normalized by parts by the labels. The outcome of the normalization looks much better (see Figure 37). But the problem with this normalization is that blots rarely contain whole set of the labels. For example only the second strip in Figure 37 has the third label. And while the strips in Figure 37 still have quite a lot of labels there are also strips with just three labels – the zeroth, the first and the second label. In these cases the normalization is effective only on the left part of the strip. And the right part of the strip remains unaligned.

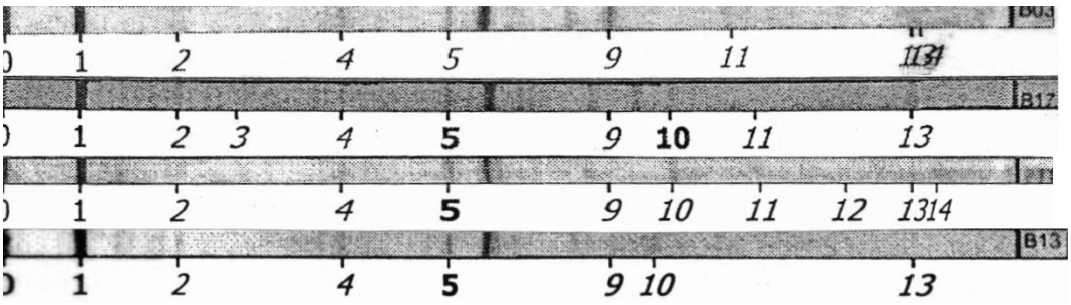

Figure 37: The strips are normalized by parts by the labels. Notice some labels are missing.

Hence the last remaining chance how to align the bands is to normalize the strips by the position of the distinctive bands. And this kind of the problem is solvable by Dynamic Time Warping (DTW).

The example output of unconstrained DTW implemented by Pau Micó (Micó, 2010) is in Figure 38. While the bands were correctly aligned to the bands in the first strip, some parts of the strips were stretched too much while other parts were cut off. This effect is easily visible on the labels, which are sometimes hardly readable.

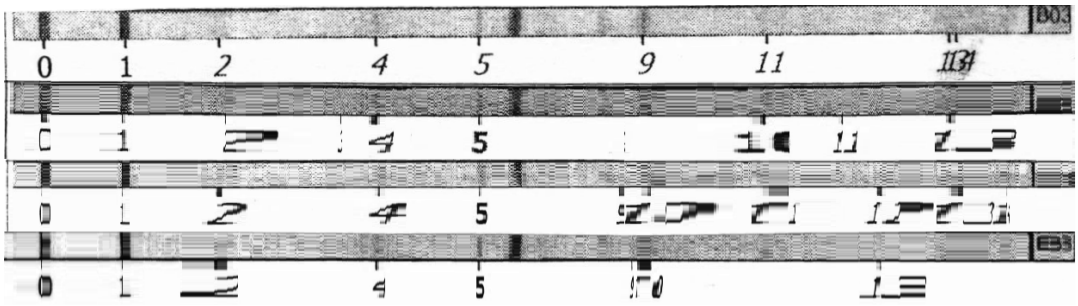

Figure 38: The strips are normalized with unconstrained Dynamic Time Warping algorithm. The bands are aligned but some columns were stretched too much (look at the labels).

The solution to the unconstrained stretching is to use constrained DTW, which limits its elasticity of the stretching. The example output with Itakura parallelogram constrain is in Figure 39. The bands are aligned approximately equally well as in the unconstrained case but the overall appearance of the normalized strips is smoother. And smoother appearance means more samples from the original strip is used in the normalized strip. Hence the constrained alignment is considered superior to the unconstrained alignment.

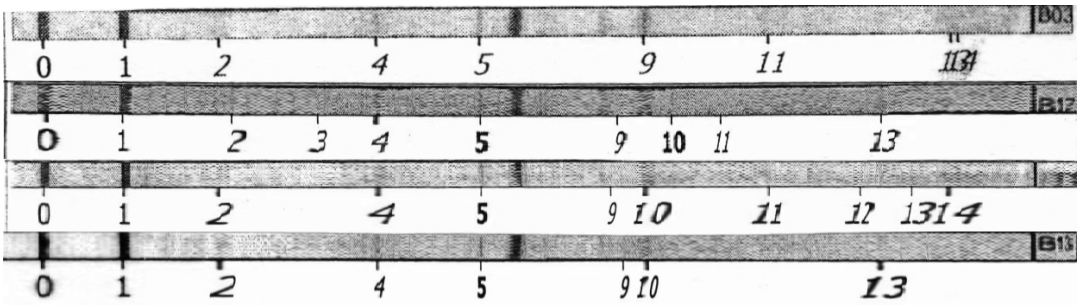

Figure 39: The strips are normalized with constrained Dynamic Time Warping algorithm.

However, some labels (for example label 11 in Figure 39) remains unaligned after DTW. This is because of the lack of the pattern in the area. And while this flaw is not a problem in the illustrated example because all the strips are from a single patient and contain basically the same bands, the situation gets complicated when we compare strips from different patients with different bands (see Figure 40).

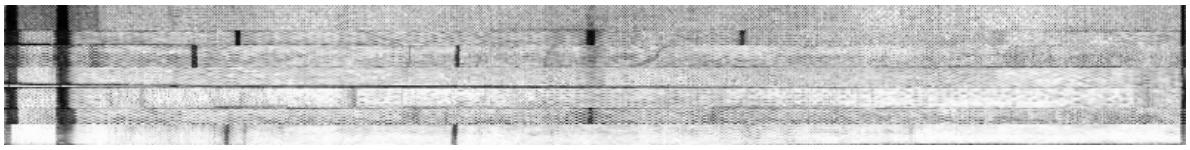

Figure 40: The picture shows strips from three patients, where each patient has two strips. The strips were aligned with DTW by the top strip. Since the top strip misses bands present in the other bands, the alignment is not perfect.

To fix this issue sample cut-off strips for IgG and IgM were obtained from the doctor and used as the reference. The resulting alignment of the same strips as in Figure 40 is in Figure 41.

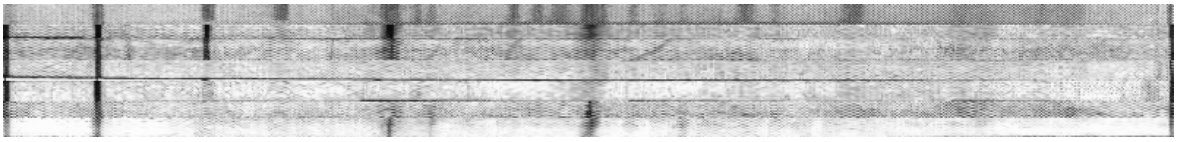

Figure 41: The picture shows strips from three patients, where each patient has two strips. The strips were aligned with DTW by an example cut-off strip (the first strip at the top).

Another step forward is to average the aligned strips within the patients.

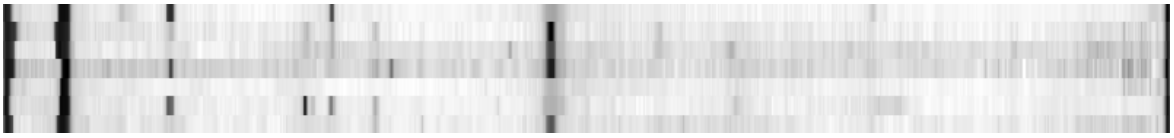

Figure 42: Average IgG strips of seven.

2.2.3 Conversion to Signal

Even though the normalized strips can be now used by machine learning algorithms, another common preprocessing step, dimension reduction, can be still performed. The information on the strips is stored in one dimension, but is represented in two dimensions to increase signal-to-noise ratio.

But once the data are digitized it is unnecessary to preserve the strips in two dimensions as the vertical dimension contains just noise. The most direct approach to dimension reduction would be the calculation of the column mean values. The outcome of this approach is in Figure 43.

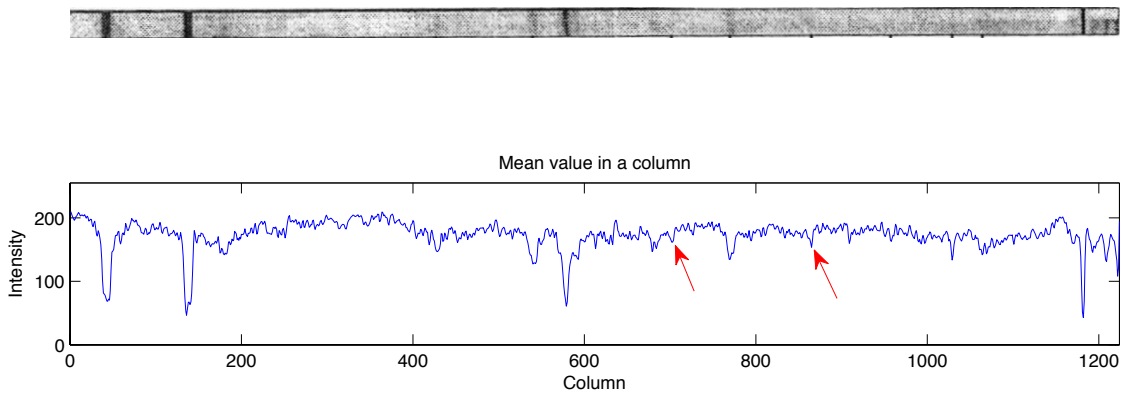

Figure 43: The strip (top). And it's mean value (bottom). Notice small peaks marked by red arrows – they are not in the blot but they were propagated into the mean value from band labels bellow the strip (small black dots bellow the top strip), which weren't perfectly removed during the stripping phase.

This direct approach is not satisfactory because the band labels from below the strip propagates into the mean value and create false bands. The solution is to remove a few bottom lines and then the mean value will work. To get the number of lines to remove we can minimize mean column variance. Ideally the values in a column are the same and the variance is zero. But if there is a black border at the top or at the bottom the variance increases. This is illustrated in the uphill part in Figure 44, which shows dependence of the mean variance in a column on the count of rows used to calculate the variance. After the 30th row from the top the rows begin to contain the black border and the variance increases.

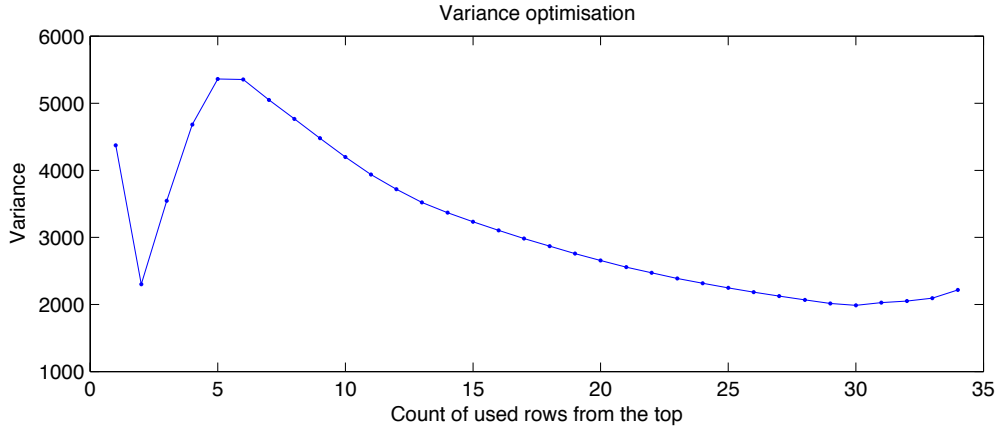

Figure 44: Dependence of the mean variance in a column on the count of used rows. The minimal variance is at the 30th row.

However, if we use less than 30 rows the variance also increases. This is because the strip has also top black border. This top biases the true column mean value. And this bias increases with decreasing count of used rows. Hence all we have to do is to find the global local minimum of the mean variance. Similarly the top black border can be removed by building the graph in Figure 44 but beginning from the bottom row. An illustration of what is removed by this method is in Figure 45.

Figure 45: Magenta color highlights rows removed is optimized by minimum mean variance criterion.

After the removal of the noisy borders we get the mean plot depicted in Figure 46. The false peaks disappeared.

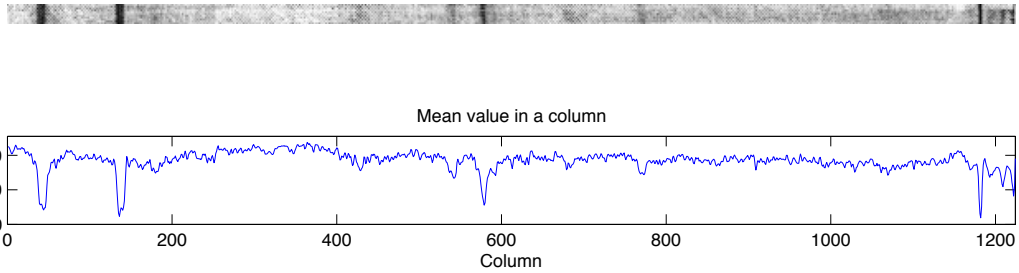

Figure 46: The mean plot of the same strip depicted in Figure 43 but after removing the border rows.

3 Experimental Results

Because data preprocessing is an important part of data mining process, the performance of the Western blot registration was examined. Then the results of the cluster analysis, which was annotated together with the domain expert, is presented.

3.1 Western Blot Registration

Two aspects of the registration were measured– accuracy and speed. The accuracy is measured by the proportion of the detected strips:

$$proportion\ of\ detected\ strips\ =\ count\ of\ found\ strips/2 * count\ of\ pages$$

Because the registration is time consuming and there are many parameters in the model to change, the experimentation was limited to changing one parameter at once from the following default setting:

Denoise filtering: median filter [3×3].
Horizontal correction: Dominant objects rotation.
Binarization: Triangle thresholding.
Morphological transformations: opening [4×4] followed by closing [7×7].

Figure 47 depicts the measured results for each tested change from the default configuration.

Changed	Method	Found [%]	Time [s]
Reference setting	Described in the text.	79.4	1,90
Workflow change	Swap morph. trans. & binariz.	**86.3**	3,73
Filtering	Without filtering	79.9	1,70
	Wiener	**83.2**	1,84
	Unsharp mask	67.2	1,90
Horizontal correction	Without correction	77.5	1,68
	Fourier transformation	**80.7**	2,47
	Hough transformation	**81.5**	2,18
Thresholding	Without opening	68.1	1,54
	Otsu's method	27.0	2,08
Morphological opening	Opening [3×3]	79.1	1,86
	Opening [5×5]	79.2	1,91
	Opening [6×6]	78.1	1,86
	Opening [7×7]	79.9	2,00
	Opening [10×10]	**94.4**	2,67
	Opening [13×13]	**94.2**	3,51
	Opening [17×17]	**85.0**	4,05
Morphological closing	Without closing	65.2	2,31
	Closing [6×6]	79.2	1,94
	Closing [8×8]	79.9	1,98

Closing [10×10]	80.1	2.15
Closing [12×12]	80.3	2.56
Closing [13×13]	80.4	3.60
Closing [14×14]	79.5	4.24
Closing [16×16]	79.4	4.89

Figure 47: Influences of algorithm modifications. The number of false positive is zero for all tested combinations and is omitted from the table. The time is the average time for processing one report on Mac 10.7.5 with 1.7 GHz processor and 4 GB RAM. Numbers in bold text have better accuracy than the default setting.

Based on the results in Figure 47 each block from the strip registration workflow improves accuracy (for example without the closing operation the accuracy would be lower by 15 percent points). And some changes against the default setting improve the accuracy: Wiener filtering outperforms median filtering and Hough transformation horizontal correction outperforms dominant objects rotation. Also putting the morphological operations before binarization improves the accuracy. Hence a new optimized version was developed and tested:

Filtering: Wiener filter.
Horizontal correction: Hough transformation.
Binarization: Triangle thresholding.
Morphological transformations: opening [10×10] & closing [13×13].
Order: morphological operations first, then binarization.

The result of this optimized version is compared against the original setting in Figure 48.

Method	Found [%]	Time [s]
Default	79.4	1.90
Optimized	**98.0**	5.76

Figure 48: Comparison of the initial default method with the optimized method.

3.2 Clustering

Based on exhaustive comparison of different linkage methods (single, complete and average), different metrics (Euclidean, Manhattan and correlation) and signal normalization methods (raw data, z-score normalization, quantile normalization) the best performing combination in hierarchical clustering is z-score normalized data, correlation metric and average linkage. This combination is popular among bioinformatics for microarray assessment (Motl, Knowledge-Oriented Molecular Classifiers, 2010) and Western blots (BIO-RAD, 2013) and hence it is not surprising it works well (Figure 49).

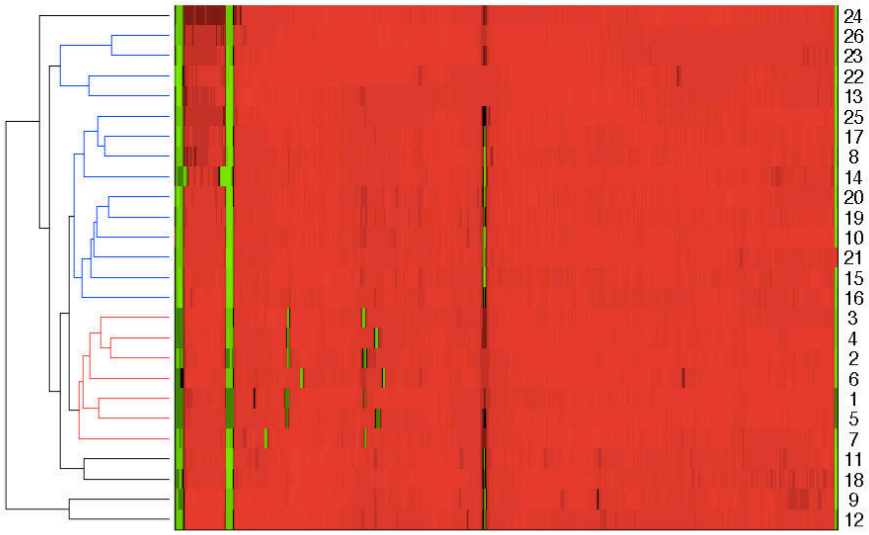

Figure 49: Clustergram of 26 IgG strips from four patients {1,2,3,4,5,6,7}, {8}, {9,10,11,12} and {13,14,15,16,17,18,19,20,21,22,23,24,25,26} with Dynamic Time Warping metric. The red branch marks the 1st patient. The blue branch marks the 4th patient.

A bigger example of IgG Western blots clustered with Dynamic Time Warping is in Figure 50. The numbers on the right shows laboratory finding on IgG Western blots. Since positive patients were grouped together we may conclude that clustering works. And consequently that preprocessing steps are sufficient.

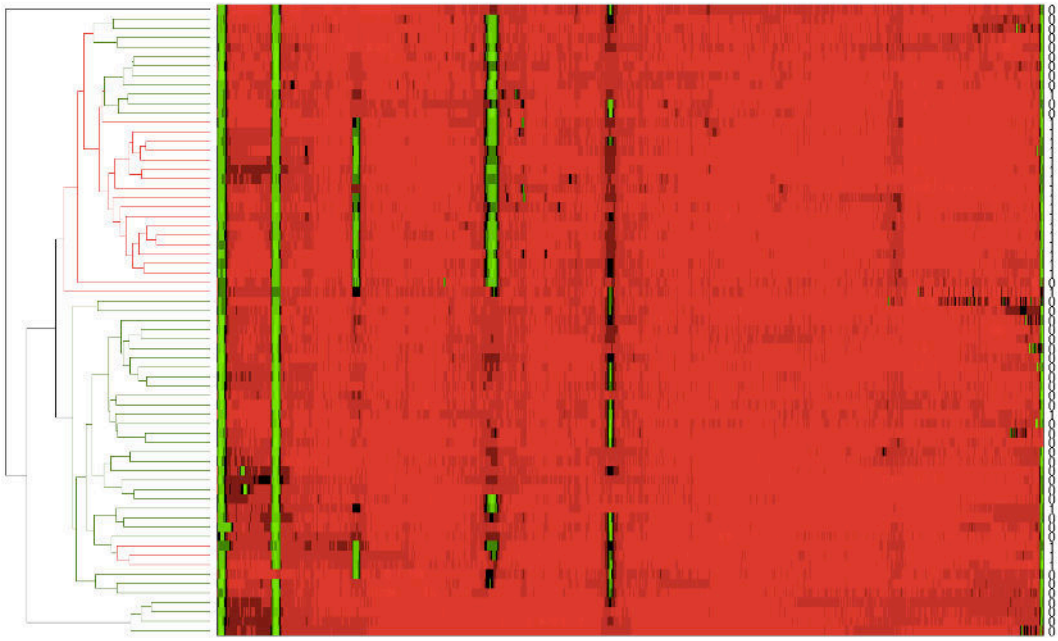

Figure 50: Clustergram of 71 IgG strips with Dynamic Time Warping metric. The labels on the right mark positive (1) or negative (0) laboratory finding on IgG strips.

Because hierarchical clustering gets quickly cluttered with the increasing number of the used samples, a Self Organizing map was produced. Figure 51 displays all IgG

strips in the map. The blue dots represent Lyme disease negative patients and red dots represent Lyme disease positive patients.

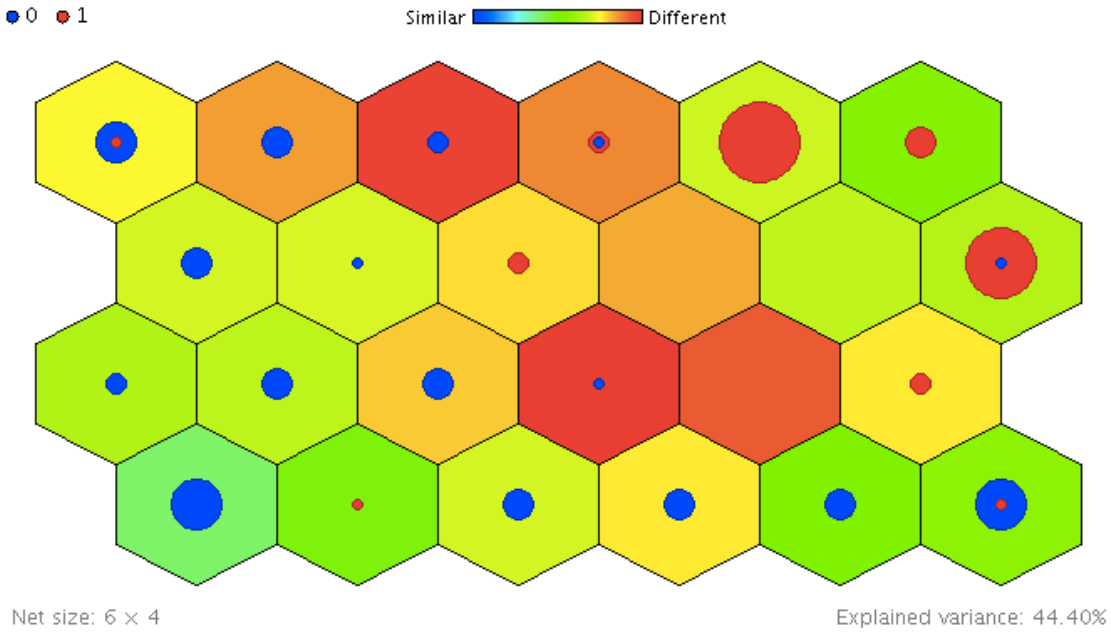

Figure 51: Self Organizing Map of IgG strips.

Now we look at two particular patients. Both of them were treated with antibiotics, one because of Lyme disease and the second because of a different illness. Still both patients display the same movement on the map. Once they start taking the antibiotics they move right down on the map. And once they stop taking them they move left up on the map (Figure 52).

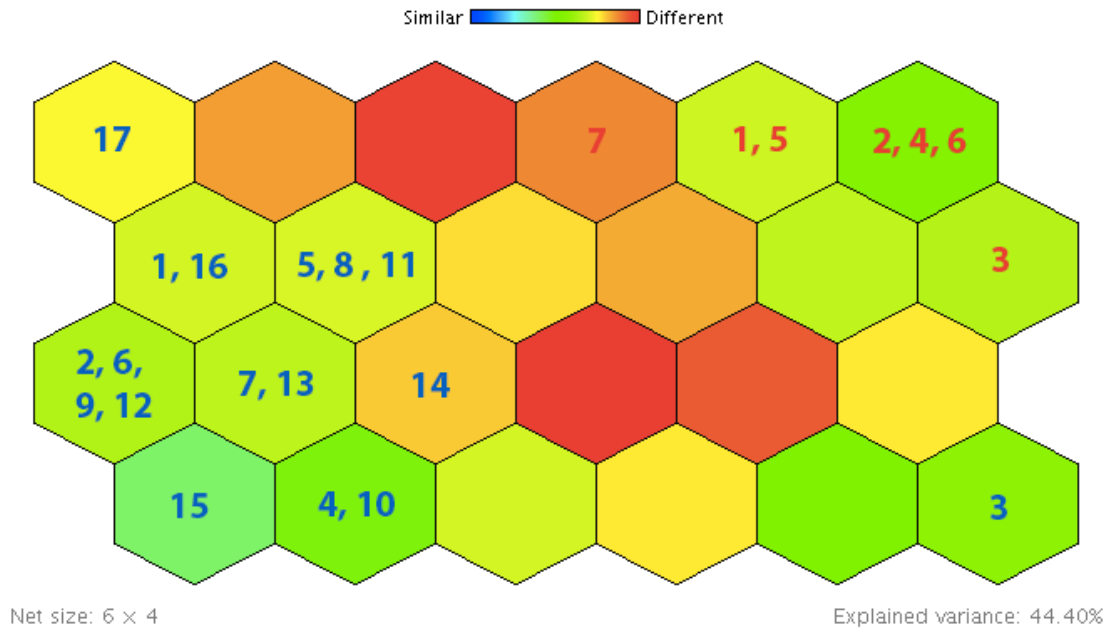

Figure 52: Evolution of two patients in SOM. The Lyme positive patient is in red color and the Lyme negative patient is in blue color.

Map with 12 patient's paths are displayed in Figure 53 and Figure 54.

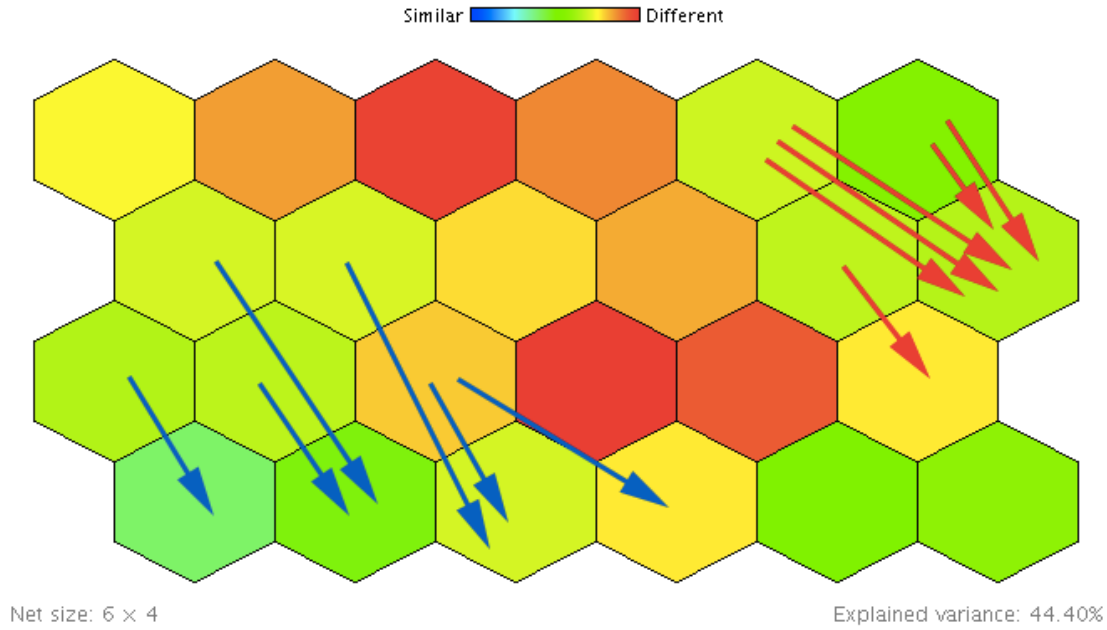

Figure 53: Path of 12 patients once they start using antibiotics.

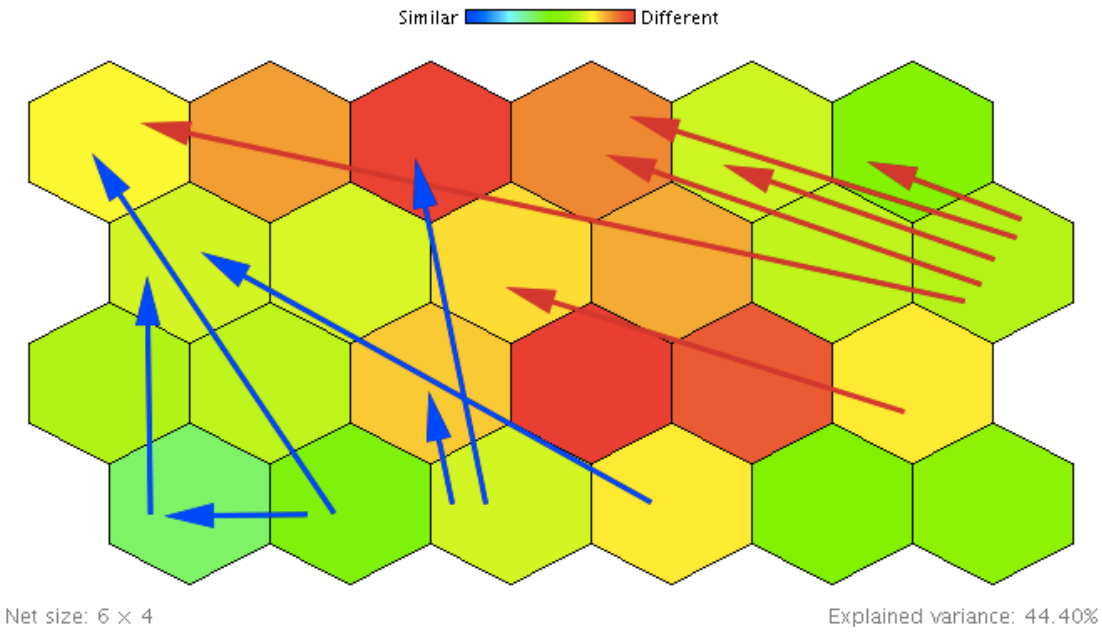

Figure 54: Path of the 12 patients once they stop using antibiotics.

4 Comparative Study of Designs of Similar Systems

It's hard to evaluate the developed system when there is not any benchmark or system to compare. However, based on the screening of tools used by scientists to analyze Western blots performed by National Institute of Standards and Technology (Bright, 2011) popular tools are: ImageJ, ImageQuant TL (GE Healthcare, 2013), Quantity One 1-D Analysis Software (BIO-RAD, 2013) and Image Studio Lite (LI-COR, 2013). ImageJ is a general image processing tool extensible with Java plugins. ImageJ allows great flexibility and possibilities with image registration plugin and comfortable GUI. However, it misses capabilities like band aligning and automatic horizontal straightening. ImageQuant and Quantity One are examples of commercial products bundled together with Western blot machines. Because of that it wasn't possible to test them. But based on product description ImageQuant provides:

> *Its high level of automation throughout allows for fully automatic analysis of 1-D gels - including lane creation, background subtraction, band detection, molecular weight calibration, quantity calibration, and normalization - in a matter of a few seconds.*

On the other hand Quantity One is described with following text:

> *Automatic lane and band detection*
>
> *Rapid molecular weight determination with choice of multiple regression models and preset standards*
>
> *Band and lane matching analysis with comparative dendrogram creation*
>
> *Background subtraction correction of gradient gels*
>
> *Purity analyzer*

Again, both these products provide automatic Western blot registration (lane detection). Additionally they provide molecular weight and quantity calibration for their company blots. Nevertheless, this calibration is likely bounded to additional data from their machines (cut-off strips…). Another aspect of Quantity One is that they also use hierarchical clustering to match blots.

And finally comes Image Studio, which is a free tool for Western blot analysis. Unfortunately, Image Studio is just a specialized version ImageJ, which doesn't provide more then ImageJ but better annotation capabilities.

5 Recommendation for Further Work

Based on the data mining workflow in Figure 8 the future work should continue where the thesis ended. The thesis findings should be compared with biomedical resources and validated. Once the findings are validated a web application will be designed and published on a web dedicated to Lyme disease (Borelioza CZ, 2012) with 80 000 visits a month as dealt with the domain owner, Ing. Petr Dymáček, Ph.D. The application will be designed for Lyme disease patients and not for doctors because experiments tells us that doctors tends to dismiss machine learning products despite their superiority (Goldberg, 1968). On the other end patients may be interested in using the application because the application is going to provide different information about their disease than their doctors can provide.

The core functionality of the application is going to be based on a history of Western Blot strips. So far each evaluation of Lyme disease is based on a single report, in better case on two consequent reports. But computer is capable evaluating whole patient's history. The patient makes a photo of the strip with his/her cellphone, uploads the strip, fill in the date and the strip will be put into the patient's history. And based on the strips in the history the patient's progress will be marked on an annotated map. And not only that the map will show how long did the patient had Lyme disease but also a prediction how much time will it take to completely recover. We can also identify the strain of Borrelia that infected the patient and so on. Making the web application, popularizing it and providing support to the application would finish the project's life based on the data mining workflow in Figure 8.

A potential sidestep to the work done in the thesis is using patient's symptoms to analyze patient's Lyme disease. A general-purpose web dedicated to diagnosis patient's problem based on the symptoms can be found for example at WebMD (WebMD, LLC., 2013). And their web likely works since they are the 408th most visited domain in the world (Alexa Internet, Inc., 2013). However, WebMD can't diagnose illnesses down to the details. And Lyme disease is a perfect candidate for this drill down. There is not any reliable test to diagnose Lyme disease, thus symptoms are used by doctors as an important lead. Lyme disease has a plenty of symptoms (at least 90 somatic and another 20 psychosomatic (Dymacek, 2012)). And each Borrelia strain is known to cause a bit different symptoms (Franke, 2013). The best thing on this sidestep is that the data are already available. The web dedicated to Lyme disease (Borelioza CZ, 2012) has been collecting Lyme disease symptoms from visitors and so far they collected data from over 400 people. Unfortunately they don't use the collected data for anything else than a general statistics. Hence there is a great potential for personalized approach with machine learning methods.

Combining test results with symptoms is promising to provide more accurate and complete diagnosis than either data itself.

6 Contribution

6.1 Dataset

The dataset with the Western blot signals and other metadata was published at Machine Learning Repository (UCI, 2013) where anyone can download them and reexamine them. The dataset composes of following attributes: normalized signal of IgG, normalized signal of IgM, patient ID, date of taking blood sample, laboratory Western blot finding and laboratory ELISA finding. Detail description of the dataset, which follows the dataset exchange format (UCI, Documenting Requirements, 2013), is in the appendix C.

6.2 Code

The reusable code produced during the work on the thesis was uploaded to Mathworks File Exchange (Mathworks, 2013). The uploaded functions include automatic image straightening, fast mean filter based on integral image that is faster than MATLAB's mean filter in Image Processing Toolbox (Figure 55), fast standard deviation filter that is faster than MATLAB's standard deviation filter in Image Processing Toolbox (Figure 55), implementation of different thresholding methods (Bradley, Bernsen, Feng, Niblack, Sauvola) that are using the fast mean and standard deviation filters (the example outputs are in Appendix E), function for rotation of an image about any point in the image (Image Processing Toolbox includes only a method for image rotation about the image's center), Magic kernel edge detection and function that returns Sobel filter of arbitrary size. The total download rate is above 400 downloads a month (Figure 57).

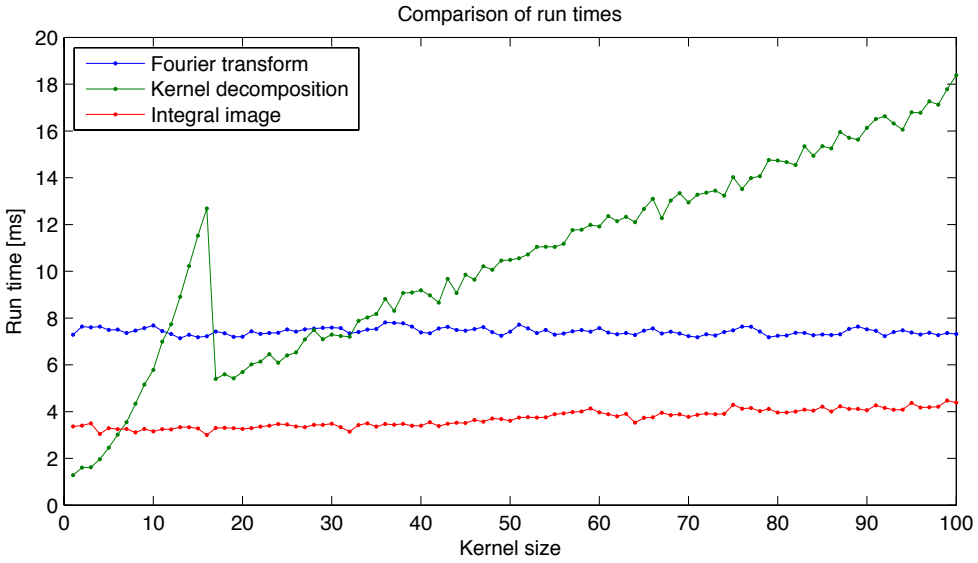

Figure 55: Comparison of run times based on the kernel size for Fourier transformation (MATLAB 2011a), kernel decomposition (*fspecial* function from Image Processing Toolbox) and integral image. The run time for the integral image is slightly increasing because of activated border deformation correction.

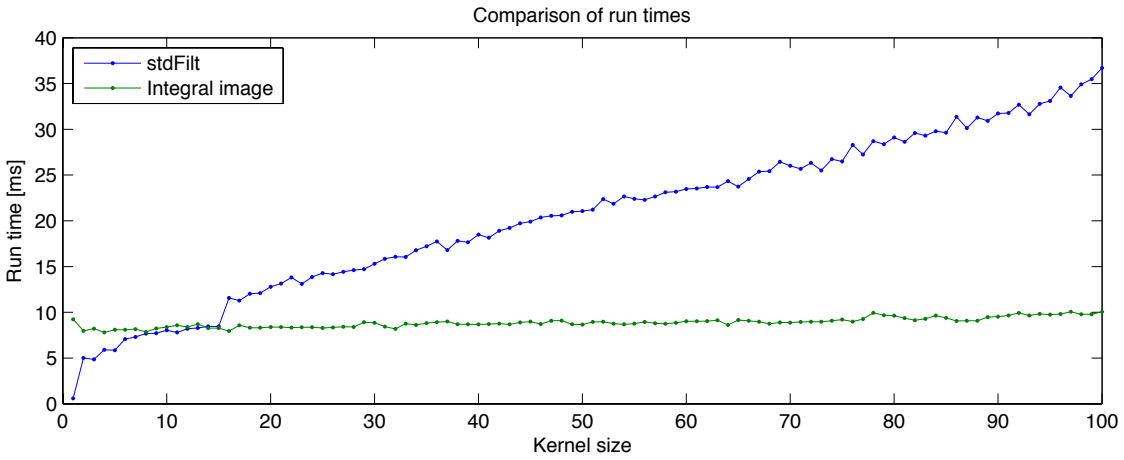

Figure 56: Comparison of run times based on the kernel size for *stdfilt* (from Image Processing Toolbox) and integral image. The run time for the integral image is slightly increasing because of activated border deformation correction.

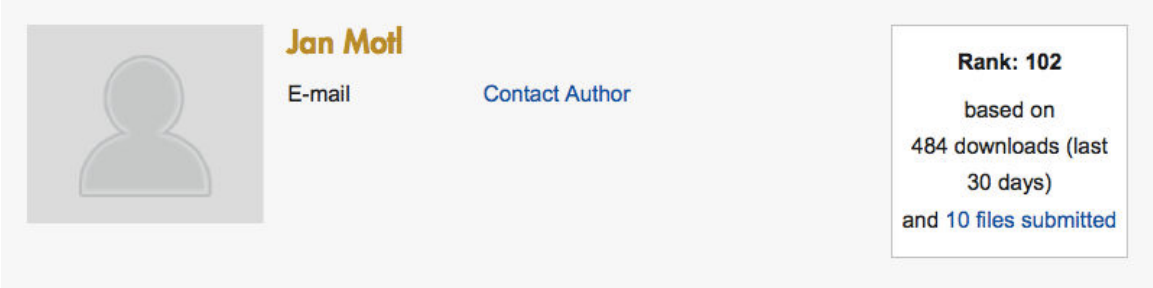

Figure 57: Count of downloads for reusable functions developed for this thesis.

Additionally a SOM plugin for RapidMiner was developed and uploaded to Rapid Marketplace (Rapid-I Inc., 2012). The plugin has over 6000 downloads

Platform	Version	Release date	File size	License
ANY	5.2.0	12/28/12 2:02 PM	56 kB	AGPL

Download statistics
Total: 6634; This week: 583; Today: 83
3 times bookmarked

Figure 58: Download statistics for SOM plugin.

6.3 Automated Western Blot Processing

The Western blot registration and preprocessing is a difficult problem comparable to microarray processing. Still the Western blot registration and preprocessing were implemented and fully automated including details like the normalization by labels.

7 Conclusion

After collecting, preprocessing and analyzing 4227 reports two groups of patients were identified on the SOM map. The patients with Lyme disease were on West and the patients without Lyme disease were on East. However, some Western blots were misplaced on the map. The majority of these displacements were because of imperfect preprocessing. But systematic misplacements have to be further discussed with the patient's doctor since some medical conditions, which weren't included in the analysis (*immunodeficiency*, *Helicobacter pylori*…), are known to interact with Lyme disease Western blots. Including these factors could further increase the map's accuracy.

Typical patient's movement on the SOM map was described. Once the patient begins to use antibiotics the patient's position on the map moves to Southeast. And after finishing the antibiotic treatment either the patient moves to the original position and the illness returns or the patient's location moves to Northwest of the patient's original position. This Southeast movement is known as Jarisch-Herxheimer reaction. As the dead bacteri bodies get into the blood stream the immune system strongly reacts. And this strong reaction is measured on the Western blots.

Bibliography

Choi, J. (2007). *Realtime On-Road Vehicle Detection with Optical Flows and Haar-like feature detectors.* Retrieved from CiteSeer: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.93.5581

Cytochem, H. (1977). Automatic Measurement of Sister Chromatid Exchange Frequency. *The Journal of Histochemistry and Cytochemistry* , 741-753.

Óscar Marbán, G. M. (2009). *A Data Mining & Knowledge Discovery Process Model.* Retrieved from In Data Mining and Knowledge Discovery in Real Life Applications: http://cdn.intechopen.com/pdfs/5937/InTech-A_data_mining_amp_knowledge_discovery_process_model.pdf

Alexa Internet, Inc. (2013). *webmd.com.* Retrieved from Alexa: http://www.alexa.com/siteinfo/webmd.com

Assous, M. (1993). *Western blot analysis of sera from Lyme borreliosis patients according to the genomic species of the Borrelia strains used as antigens.* Retrieved from PubMed: https://www.ncbi.nlm.nih.gov/pubmed/8513814

Aucott, J. (2012). *Bull's-Eye and Nontarget Skin Lesions of Lyme Disease: An Internet Survey of Identification of Erythema Migrans*. Retrieved from PubMed: http://www.ncbi.nlm.nih.gov/pubmed/23133445

Basile, T. (2010, October). A Texture-Based Image Processing Approach for the Description of Human Oocyte Cytoplasm . *Instrumentation and Measurement, IEEE* , 2591 - 2601 .

BIO-RAD. (2013). *Image Analysis Software.* Retrieved from BIO-RAD: http://www.bio-rad.com/prd/pt/PT/LSR/PDP/1de9eb3a-1eb5-4edb-82d2-68b91bf360fb/Quantity-One-1-D-Analysis-Software

Bogen, C. E. (1994). *Proceedings of the Socond National Conference on Serologic Diagnosis of Lyme Disease*. Retrieved from Viralab: http://www.viralab.us/Dearborn-2nd%20Nat.%20Conf.%20on%20Serologic%20Diagnosis%20of%20Lyme%20Recommendations.pdf

Bright, D. (2011). *Tools.* Retrieved from National Institute of Standards and Technology: http://www.nist.gov/lispix/doc/

Brorson, Ø. (2009). *Borrelia burgdorferi – en unik bakterie*. Retrieved from Tidsskrift for Den norske legeforening: http://dx.doi.org/10.4045/tidsskr.08.0023

Buades, A. (2013). *Implementation of the "non-local Bayes" image denoising algorithm*. Retrieved from Image Processing On Line: http://www.ipol.im/pub/pre/16/preprint.pdf

de la Escalera, A. (1997, December). *Road traffic sign detection and classification*. Retrieved from IEEE: http://dx.doi.org/10.1109/41.649946

Dymacek, P. (2012). *Borelioza CZ*. Retrieved from http://http://www.borelioza.cz

Evans, R. (2010). *More specific bands in the IgG western blot in sera from Scottish patients with suspected Lyme borreliosis.* Retrieved from PubMed: https://www.ncbi.nlm.nih.gov/pubmed/20595179

Faradji, F. (2007). A Morphological-Based License Plate Location. *Image Processing* , 57-60.

Fisher, R. (2003). *Hough Transform*. Retrieved 2013, from Image processing learning resources: http://homepages.inf.ed.ac.uk/rbf/HIPR2/hough.htm

Franke, J. (2013, February). *Exploring gaps in our knowledge on Lyme borreliosis spirochaetes – Updates on complex heterogeneity, ecology, and pathogenicity*. Retrieved from http://dx.doi.org/10.1016/j.ttbdis.2012.06.007

Garaiová, M. (1990). Lyme borreliosis - the big imitator. *Československá Patologie , 26*, 112-118.

GE Healthcare. (2013). *ImageQuant.* Retrieved from GE Healthcare: http://www.gelifesciences.com/webapp/wcs/stores/servlet/catalog/en/GELifeSciences-CZ/products/AlternativeProductStructure_16016/

Goldberg, L. (1968). Simple models or simple processes? *American Psychologist , 23* (7).

Goossens, H. (1999). *Evaluation of fifteen commercially available serological tests for diagnosis of Lyme borreliosis.* Retrieved from PubMed: https://www.ncbi.nlm.nih.gov/pubmed/10517192

GSA. (2010). *GSA Image Analyser*. From GSA: http://image.analyser.gsa-online.de/

Hassanzadeh, S. (2011). *Fast logo detection based on morphological features in document images*. Retrieved from http://dx.doi.org/10.1109/CSPA.2011.5759888

Hauser, U. (1998). *Diagnostic value of proteins of three Borrelia species (Borrelia burgdorferi sensu lato) and implications for development and use of recombinant antigens for serodiagnosis of Lyme borreliosis in Europe.* Retrieved from PubMed: https://www.ncbi.nlm.nih.gov/pubmed/9665948

Huegli, D. (2011). *Prospective study on the incidence of infection by Borrelia burgdorferi sensu lato after a tick bite in a highly endemic area of Switzerland.* Retrieved from PubMed: https://www.ncbi.nlm.nih.gov/pubmed/21890065

Itakura, F. (1975). *Minimum prediction residual principle applied to speech recognition*. Retrieved from IEEE: http://dx.doi.org/10.1109/TASSP.1975.1162641

Jang, R. (2012). *Machine Learning Toolbox*. Retrieved from Dynamic Time Warping: http://mirlab.org/jang/matlab/toolbox/machineLearning/help/dtw_help.html

Jung, C. R. (2004). *Rectangle Detection based on aWindowed Hough Transform.* Retrieved from IEEE: http://doi.ieeecomputersociety.org/10.1109/SIBGRA.2004.1352951

Kaplan, M. (2013). *Interpreting the IgG & IgM Western Blot For Lyme Disease*. Retrieved from Herp Care Collection: http://www.anapsid.org/lyme/wb.html

Karlsson, M. (1989). *Comparison of Western blot and enzyme-linked immunosorbent assay for diagnosis of Lyme borreliosis.* Retrieved from PubMed: https://www.ncbi.nlm.nih.gov/pubmed/2512131

Keogh, E. (2013). *Eamonn Keogh.* Retrieved from University of California - Riverside: http://www.cs.ucr.edu/~eamonn/

Landini, G. (2013). *Fiji Is Just ImageJ*. Retrieved from Auto Threshold: http://fiji.sc/wiki/index.php/Auto_Threshold

LI-COR. (2013). *Image Studio Lite.* Retrieved from LI-COR: http://www.licor.com/bio/products/software/image_studio_lite/index.jsp

Lyme disease org. (2012). *Lyme Disease Diagnosis*. Retrieved from Lyme Disease: http://www.lymedisease.org/lyme101/lyme_disease/lyme_diagnosis.html

Mahmood, N. H. (2011). Comparison between Median, Unsharp and Wiener filter and its effect on ultrasound stomach tissue image segmentation for Pyloric Stenosis. *International Journal of Applied Science and Technology , 1* (5).

Marangoni, A. (2005). *Comparative evaluation of two enzyme linked immunosorbent assay methods and three Western Blot methods for the diagnosis of culture-confirmed early Lyme borreliosis in Italy.* Retrieved from PubMed: https://www.ncbi.nlm.nih.gov/pubmed/15782625

Mathworks. (2013). *Jan Motl.* Retrieved from Mathworks File Exchange: http://www.mathworks.com/matlabcentral/fileexchange/authors/303241

Mavin, S. (2011). *Interpretation criteria in Western blot diagnosis of Lyme borreliosis.* Retrieved from PubMed: https://www.ncbi.nlm.nih.gov/pubmed/21473255

Mavin, S., & Evans, R. (2009). *Local Borrelia burgdorferi sensu stricto and Borrelia afzelii strains in a single mixed antigen improves western blot sensitivity*. Retrieved from PubMed: https://www.ncbi.nlm.nih.gov/pubmed/19240047

Micó, P. (2010). *Dynamic Time Warping*. Retrieved from Matlab Central: http://www.mathworks.com/matlabcentral/fileexchange/16350-continuous-dynamic-time-warping/content/dtw.m

Motl, J. (2011). *Cell counting.* Prague: Semestral work for MI-PDD at FIT, CTU.

Motl, J. (2010). *Knowledge-Oriented Molecular Classifiers.* Prague: Czech Technical University.

National Institute of Allergy and Infectious Diseases. (2012). *Lyme Disease*. Retrieved from NIH: http://www.niaid.nih.gov/topics/lymedisease/research/pages/diagnostics.aspx

Panneton, B. (2010). *Gray image thresholding using the Triangle Method*. Retrieved from Matlab Central: http://www.mathworks.com/matlabcentral/fileexchange/28047-gray-image-thresholding-using-the-triangle-method

Pavia, C. (1998). An Understanding of Laboratory Testing for Lyme Disease. *Journal of Spirochetal and Tick-borne Disease , 5*.

Rapid-I Inc. (2012). *Self Organizing Map.* Retrieved from Rapid Marketplace: http://rapidupdate.de:8180/UpdateServer/faces/product_details.xhtml?productId=rmx_som

Ratanamahatana, C. A. (2004). *Three Myths about Dynamic Time Warping Data Mining*. Retrieved from University of California: http://www.cs.ucr.edu/~ratana/RatanamC.pdf

StatSoft, Inc. (2009). *Cluster Analysis*. Retrieved from StatSoft: http://www.statsoft.com/textbook/cluster-analysis/

The MathWorks, Inc. (2009). *Create agglomerative hierarchical cluster tree*. Retrieved from MathWorks: http://www.mathworks.com/access/helpdesk/help/toolbox/stats/linkage.html

Tylewska-Wierzbanowska, S. (2002). *Limitation of serological testing for Lyme borreliosis: evaluation of ELISA and western blot in comparison with PCR and culture methods.* Retrieved from PubMed: https://www.ncbi.nlm.nih.gov/pubmed/12422608

UCI. (2013). *Documenting Requirements*. Retrieved from Center for Machine Learning and Intelligent Systems: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/DOC-REQUIREMENTS

UCI. (2013). *UCI*. Retrieved from Center for Machine Learning and Intelligent Systems: http://archive.ics.uci.edu/ml/

Valtameri. (2011). *Lymeská borrelióza*. Retrieved from Wiki Skripta: http://www.wikiskripta.eu/index.php/Lymesk%C3%A1_borreli%C3%B3za

Viola, P., & Jones, M. (2001). *Rapid Object Detection using a Boosted Cascade of Simple Features.* Retrieved from Microsoft: http://research.microsoft.com/~viola/Pubs/Detect/violaJones_CVPR2001.pdf

WebMD, LLC. (2013). *WebMD.* Retrieved from http://http://symptoms.webmd.com

Wikimedia Foundation, Inc. (2013). *Cross Industry Standard Process for Data Mining.* Retrieved from Wikipedia.

Abbreviations and Terms

Borreliosis	A different name for Lyme disease.
Borrelia	A bacteria responsible for Lyme disease.
DTW	Dynamic Time Warping is an algorithm for aligning signals.
ELISA	Enzyme-Linked ImmunoSorbent Assay, a tool for Lyme diagnosis.
IgG	Immunoglobulin G, a protein responsible for immunity.
IgM	Immunoglobulin M, a big protein responsible for immunity.
WB	Western blot, a tool used in Lyme disease diagnosis.

Appendix A: List of Papers about Lyme Disease

Link	# WB	Year
Western blot analysis of sera from Lyme... (Assous, 1993)	52	1993
Comparative evaluation of two enzyme... (Marangoni, 2005)	60	2005
Comparison of Western blot and enzyme-linked... (Karlsson, 1989)	69	1989
Limitation of serological test... (Tylewska-Wierzbanowska, 2002)	90	2002
Prospective study on the incidence... (Huegli, 2011)	186	2011
Evaluation of fifteen commercially... (Goossens, 1999)	229	1999
Local Borrelia burgdorferi sensu stricto... (Mavin & Evans, 2009)	233	2009
Diagnostic value of proteins of three… (Hauser, 1998)	330	1998
More specific bands in the IgG western blot... (Evans, 2010)	511	2010
Interpretation criteria in Western blot... (Mavin, 2011)	832	2011

Appendix B: The Letter

ÚŘAD PRO OCHRANU OSOBNÍCH ÚDAJŮ

UOOUX0056K6U

Pplk. Sochora 27, 170 00 Praha 7

tel.: 234 665 555, fax: 234 665 444

e-mail: posta@uoou.cz, www.uoou.cz

Čj. UOOU-00831/13-2

Vyřizuje: JUDr. Jiří Žůrek

Vážený pan

Jan Motl

[motljan1@fit.cvut.cz]

Praha 31. ledna 2013

Vážený pane,

k Vašemu podání obdrženému dne 28. ledna 2013 Vám sdělujeme:

Pokud budete v rámci diplomové práce publikovat výsledky testu boreliózy dle Western Blotu, aniž by bylo možné výsledky testu přiřadit ke konkrétní osobě, tj. nebude v diplomové práci u jednotlivých výsledků uváděno rodné číslo, ale jak uvádíte, bude uveden jiný „náhodný" identifikátor, který budete znát pouze Vy a lékař, bylo by možné tímto způsobem údaje využít pro vědecké účely, resp. pro diplomovou práci. Zákon č. 101/2000 Sb., o ochraně osobních údajů a o změně některých zákonů, v platném znění, v jeho § 5 odst. 1 písm. e) stanovuje, že při použití osobních údajů pro vědecké účely je třeba dbát práva na ochranu před neoprávněným zasahováním do soukromého a osobního života subjektu údajů a osobní údaje anonymizovat, jakmile je to možné. Ve Vámi uvedeném případě není získání souhlasu pacientů bezpodmínečnou podmínkou pro uskutečnění Vašeho záměru.

S pozdravem

Mgr. Ladislav Hejlík v. r.
vedoucí oddělení stížností a konzultací

Appendix C: Databases Documentation

1. Title: Signals from Lyme disease Western blots

2. Sources:

   Motl, Jan. Supporting the Diagnosis of Borreliosis by Machine Learning Methods. Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2013.

3. Past Usage:

   a) The data set was used a thesis: Supporting the Diagnosis of Borreliosis by Machine Learning Methods by Jan Motl, May 2013.

   b) Indication of what attribute(s) were being predicted: The intention is to predict whether the patient is Lyme disease positive or negative.

4. Number of Instances: 4227

5. Number of Attributes:

   1605 attributes in total.

   1600 numerical attributes in range from 0 to 255, which depicts the blot's average intensity from left to right. The attributes were scaled to the uniform length of 1600 with bicubic interpolation.

   pacientid: hexadecimal pacient's id.

   date: date of accepting the blood sample for the analysis.

   elisa: a binary attribute, true if ELISA test is positive.

   wbigg: laboratory conclusion for Western blot IgG.

   wbigm: laboratory conclusion for Western blot IgM.

6. Missing Attribute Values: none

Appendix D: Used Tools

MATLAB 2011a	Programming language for rapid development.
Image Processing Toolbox	Image processing extension for MATLAB.
Bioinformatics Toolbox	Clustering in MATLAB.
Minitab	Statistics tool for data analysis.
Dynamic Time Warping	Unconstrained implementation (Micó, 2010).
Dynamic Time Warping	Constrained implementation (Jang, 2012).
Triangle Threshold	Image binarization technique (Panneton, 2010).
ImageJ	Java-based image processing toolkit.

Appendix E: Implemented Thresholding Methods

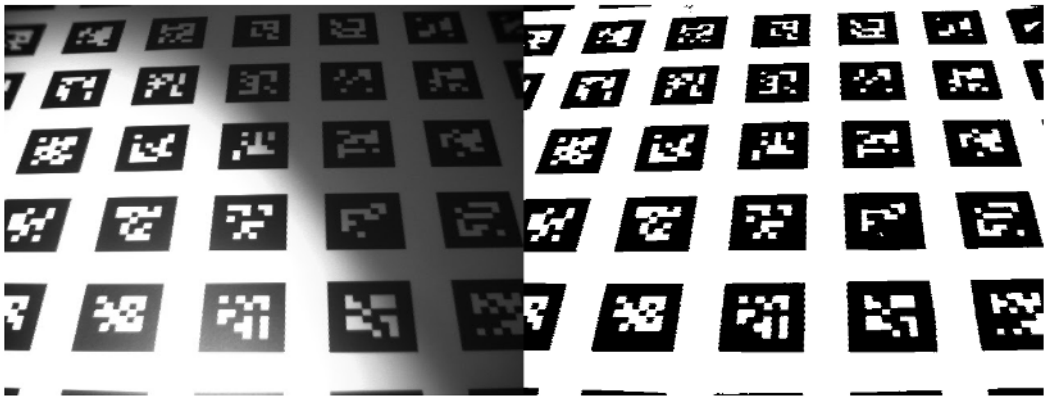

Figure 59: Niblack binarization.

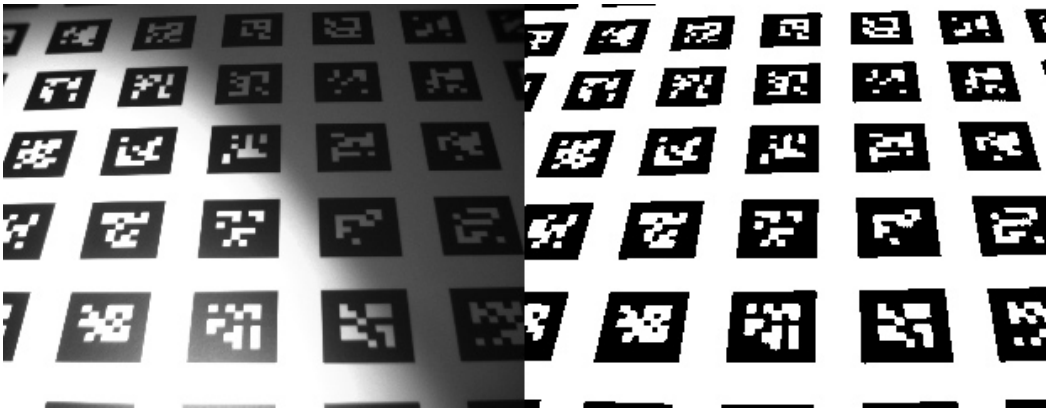

Figure 60: Sauvola binarization.

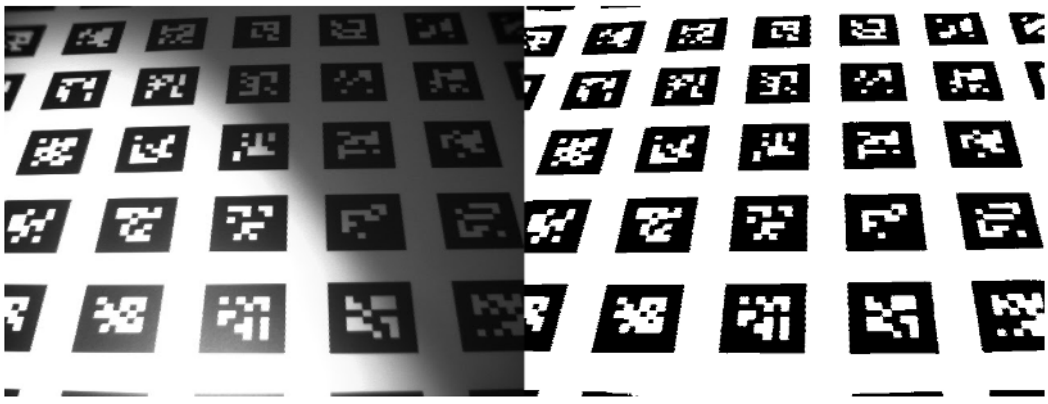

Figure 61: Bradley binarization.

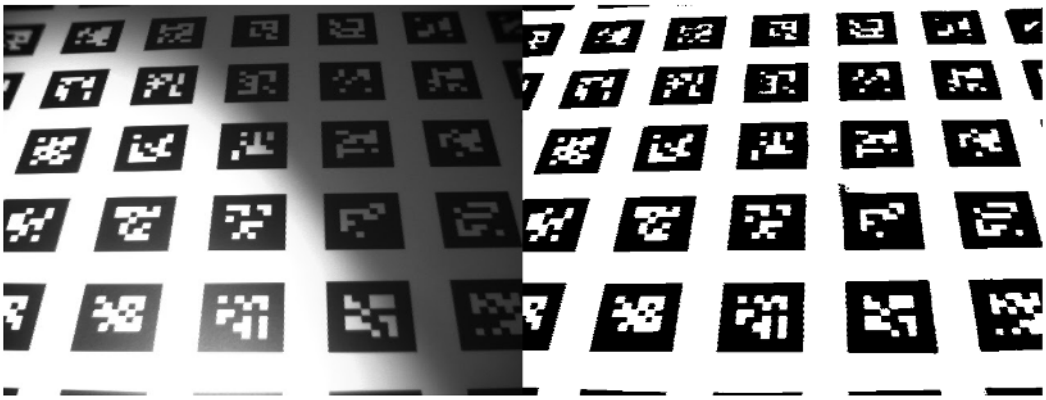

Figure 62: Bernsen binarization.

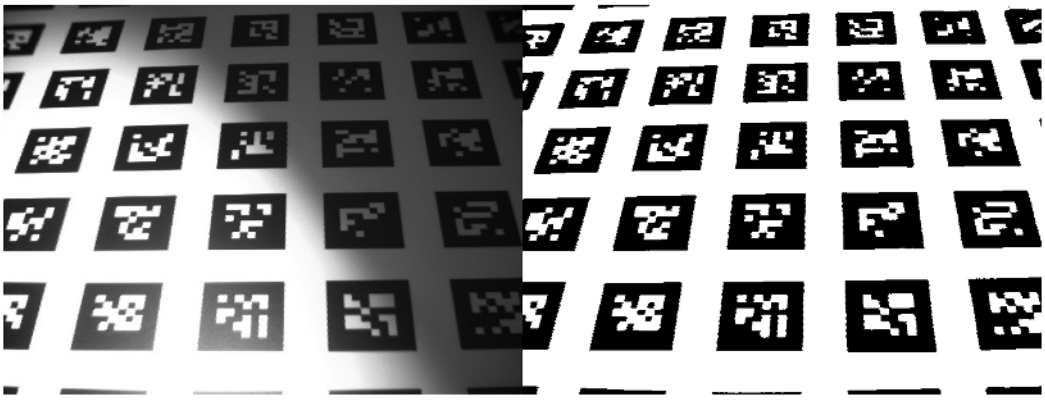

Figure 63: Feng binarization.

Appendix F: Contents of CD

readme.txt ..... the file with CD contents description
data ..... the data files directory
    wb.zip ..... the anonymised dataset
src ..... the directory of source codes
    lyme.zip ..... MATLAB code for
    horizontal alignment.zip ..... MATLAB code for horizontal alignment
    average filter.zip ..... MATLAB code for average filter
    local thresholding.zip ..... MATLAB code for local thresholding
    som.zip ..... SOM plugin for Rapid Miner
text ..... the thesis text directory
    thesis.pdf ..... the Diploma thesis in PDF format