VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2019 Diana Plišková

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

NOVÉ NÁDOROVÉ BIOMARKERY ODVOZENÉ Z KVANTITATIVNÍHO FÁZOVÉHO ZOBRAZOVÁNÍ BUNĚK
NOVEL CANCER BIOMARKERS DERIVED FROM QUANTITATIVE PHASE IMAGING OF BIOPSY CELLS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR
Diana Plišková

VEDOUCÍ PRÁCE
SUPERVISOR
doc. Ing. Jana Kolářová, Ph.D.

BRNO 2019

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní obor **Biomedicínská technika a bioinformatika**
Ústav biomedicínského inženýrství

Studentka: Diana Plišková ***ID:*** 195732
Ročník: 3 ***Akademický rok:*** 2018/19

NÁZEV TÉMATU:

Nové nádorové biomarkery odvozené z kvantitativního fázového zobrazování buněk

POKYNY PRO VYPRACOVÁNÍ:

Vedoucí práce: Ing. Daniel Zicha, CSc.;
1) Proveďte literární rešerši počítačových metod zpracování obrazu a analýzy dat relevantních pro měření reakcí živých nádorových buněk v tkáňové kultuře při použití časosběrného záznamu světelnou mikroskopií a zejména kvantitativním fázovým zobrazováním. 2) Navrhněte potenciální biomarkery a algoritmy pro jejich extrakci z mikroskopických dat. 3) Algoritmy implementujte v programovém prostředí Mathematica a ověřte pomocí reálných dat. 4) Proveďte diskusi získaných výsledků a vhodnosti biomarkerů pro použití při personální léčbě nádorového onemocnění.

DOPORUČENÁ LITERATURA:

[1] FRIEDMAN, Adam at al. 2015. “Precision Medicine for Cancer with Next-Generation Functional Diagnostics.” Nature Reviews Cancer 15 (12): 747–56. https://doi.org/10.1038/nrc4015.

[2] MILLIKEN, G A, and D E JOHANSON. 1992. Analysis of Messy Data Volume I: Designed Experiments. Journal of Marketing Research JMR. Vol. 22. New York and London: Chapman and Hall. https://doi.org/10.2307/3151596.

[3] ZICHA, D, and G A DUNN. 1995. “An Image-Processing System for Cell Behavior Studies in Subconfluent Cultures.” Journal Of Microscopy Oxford 179: 11–21.

[4] ZICHA, D at al. 1999. “TGFbeta1 Induces a Cell-Cycle-Dependent Increase in Motility of Epithelial Cells.” J Cell Sci 112: 447–54. http://www.ncbi.nlm.nih.gov/pubmed/9914157.

Termín zadání: 4.2.2019 ***Termín odevzdání:*** 24.5.2019

Vedoucí práce: doc. Ing. Jana Kolářová, Ph.D.
Konzultant:

prof. Ing. Ivo Provazník, Ph.D.
předseda oborové rady

ABSTRACT

The main objective of this work is the development of novel cancer biomarkers usable in personalized treatments. To understand why this issue is important, a brief description of cancer, including statistical results over the past years, is provided. The work also describes individual methods of light microscopy that can be used in cell analysis and subsequent image processing consisting of segmentation, tracking, feature extraction and classification. In this work, the main cell features, such as cell motility and shape, are presented. These features can be potential biomarkers in the treatment of cancer.

KEYWORDS

cancer, quantitative phase imaging, light microscopy, analysis of variance, image processing, principal component analysis, biomarker

ABSTRAKT

Hlavným cieľom tejto práce je vývoj nových nádorových biomarkerov využiteľných v perzonalizovaných liečbach. Pre pochopenie, prečo je táto problematika dôležitá, slúži stručný popis rakoviny obsahujúcej aj štatistické výsledky za uplynulé roky. Práca taktiež popisuje jednotlivé metódy sveteľnej mikroskopie využiteľné pri analýze buniek a aj následné spracovanie obrazov pozostávajúce zo segmentácie, trackingu, extrakcie príznakov a klasifikácie. V tejto práci sú prezentované príznaky hlavných vlastností buniek, ako je rýchlosť a tvar bunky. Práve tieto príznaky môžu byť potenciálne biomarkery pri liečbe rakoviny.

KLÍČOVÁ SLOVA

rakovina, kvantitatívne fázové zobrazovanie, sveteľná mikroskopia, analýza variancie, spracovanie obrazov, analýza hlavných komponent, biomarker

PLIŠKOVÁ, Diana. *Nové nádorové biomarkery odvodené z kvantitatívneho fázového zobrazovania buniek*. Brno, 2019, 70 p. Bachelor's Thesis. Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Biomedical Engineering. Advised by Ing. Daniel Zicha, CSc.

DECLARATION

I declare that I have written the Bachelor's Thesis titled "Nové nádorové biomarkery odvodené z kvantitatívneho fázového zobrazovania buniek" independently, under the guidance of the advisor and using exclusively the technical references and other sources of information cited in the thesis and listed in the comprehensive bibliography at the end of the thesis.

As the author I furthermore declare that, with respect to the creation of this Bachelor's Thesis, I have not infringed any copyright or violated anyone's personal and/or ownership rights. In this context, I am fully aware of the consequences of breaking Regulation § 11 of the Copyright Act No. 121/2000 Coll. of the Czech Republic, as amended, and of any breach of rights related to intellectual property or introduced within amendments to relevant Acts such as the Intellectual Property Act or the Criminal Code, Act No. 40/2009 Coll., Section 2, Head VI, Part 4.

Brno ....

author's signature

ACKNOWLEDGEMENT

I would like to thank the supervisor of the bachelor thesis Ing. Daniel Zicha, CSc. and consultant Ing. Tomáš Vičar for their professional mentoring, consultations, patience, and thoughtful suggestions concerning the bachelor's thesis.

Brno . . . . . .

. . . . . .

author's signature

Contents

List of Figures

List of Tables

Introduction

Cancer is one of the diseases from which humanity has been suffering for a long time. Some people think it's a disease of modern times, but the opposite is true. Recently it has come to the fore because other serious diseases such as pneumonia or infectious and bacterial diseases are easily curable. Cancer differs from these diseases because more than 200 types of cancer exist and its early course could be inconspicuous.

The results of the statistics described in the first chapter show that there are every year more and more cases of the cancer. Even with the significant progress in cancer treatment, the morbidity and mortality has been a major burden world-wide. Further improvement in treatment is required and there is a promising potential in personalized approach and pre-testing potential chemotherapeutic drugs with biopsy cells using suitable light microscopy techniques, which are described in the second chapter. It also describes the topic of the image processing, principal component analysis method and analysis of variance statistical method.

The third chapter consists of the practical part of this work where potential biomarkers are described showing the activity of cancer cells after administration of a particular drug.

1 Cancer

Cancer represents a group of over 200 individual diseases. Their main trait is malignant growth - unlimited reproduction within the range of a tissue, and invasion/metastasis - the spread of cancer cells to other parts of the organism.

As time went on with development in medical treatments, knowledge about cancer grew exponentially with it. Even though there is no full apprehension of cancer development at molecular and cellular levels, its overall characteristics and nature are widely known. [1]

1.1 Cell Biology of Cancer

Cancerogenesis is the term used to describe certain actions that cause a healthy, normal cell to change into cancerous. It consists of phases called initiation, promotion and progression. [1]

During cancerogenesis, cells change their behavior, mainly proliferative and motile, as a results of underlying genetic or epigenetics changes which reprogram signaling networks of cells.

Throughout the life of a human, about 10^{16} cell divisions will occur in the body and each one of these could lead to a malignant tumor if it gets out of control. Even one cell within contact of the carcinogens could develop series of mutations that could lead to forming the cancer. Types of mutations that may occur are:

- **Somatic** – Genetic changes may affect cells anywhere in the body, where they may form a tumor. This mutation remains in the individual, therefore it cannot be transmitted to its offspring.
- **Germinal** – This type of mutation occurs in genome of sperm or egg, or in the precursor of the gonads. Mutated cells are inherited from the individual to its offspring.
- **Chimerisms** – Transfer of mutant cells from one organism to another, for example through the placenta. [2][3][4]

Initiation is the first step towards cancer, which may occur in stem cells. It is the result of spontaneous cellular change or exposure to a carcinogen. Due to initiation agent, cells undergoe irreversible change, which in matter of time can be rapid. Genes responsible for transformation in human DNA are called oncogenes. Every oncogene has its version in a healthy DNA sequence called proto-oncogene. This type of gene can change due numerous mechanisms creating active oncogene. It may be possible that a lot of cells in the body go through the process of initiation

during the individual's lifetime, however these cells can either die spontaneously or stop spreading, or are neutralized by immunologic mechanism.

Promotion is a stage of carcinogenesis which ensue changes of transformed gene. It may remain harmless until its integrity is damaged resolving in changes of gene expression, selective clonal expansion, and proliferation of initiated cells. Compared to initiation stage, the promotion stage exhibits feature such as, its duration is prolonged, shows some characteristics of reversibility and in matter of visibility, neoplasm is notably visible. [1] [5]

Progression is name of the last stage of carcinogenesis. Its essential trait is the enormous increase of primary cancer mass, which can separate itself from its place of origin to other tissues and organs through the body, process called metastasis. As seen on Fig. 1.1, the cancer cell division lacks of control. Its effect of apoptosis is insufficient, damaged cells have developed insensitivity to anti-growth signals and self-sufficiency in growth signals, and they have limitless replicative potential. [1] [4] [5]

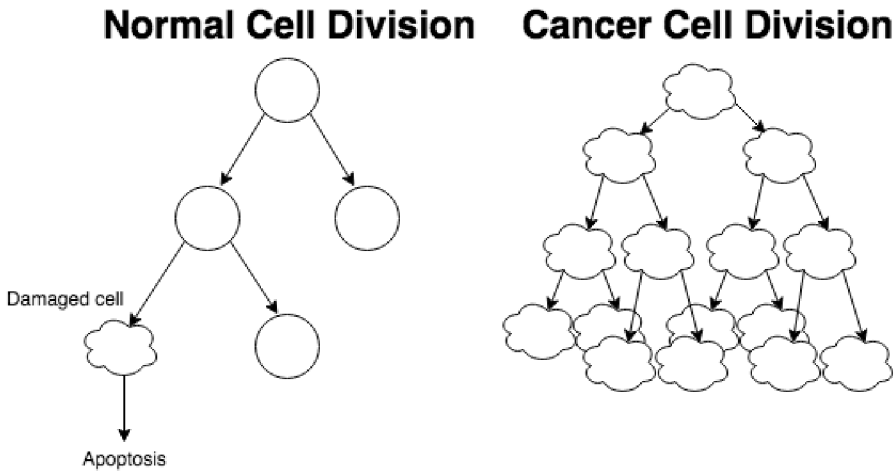


Fig. 1.1: Comparison of difference between normal and cancer cells division.

1.2 Risk Factors for Cancer

Each type of cancer has its cause, but one of the most important cures is prevention, which is more preferred option than diagnosis and following treatment. With early detection of cancer, the chance of survival increases. Epidemiologists have concluded the fact, that major causes of the cancer are in fact the results of a lifestyle of an individual. [1]

It is known that tobacco smoke is harmful not even only for a smoker, but also for non-smokers whose are within reach. It contains a lot of cancer-causing chemicals, which is why it is responsible for about 33 % of cases, leading to a lung cancer. Using smokeless products could increase risk of getting cancer of the head, neck and pancreas. In Southern Asia, chewing of tobacco-containing products is prevailing, which is the reason of major incidence of oral and pharyngeal cancer. [1] [3] [6]

Another source are numerous factors in the diet, which are mainly leading to diseases of the gastrointestinal tract, such as cancers of colon-rectum, liver or esophagus. For example, high-meat diets could be cause of previously mentioned colorectal cancer. Chemicals, such as antitoxin, byproduct of fungal intoxication in food, or nitrosamine, product of a reaction between amino acids and nitrites during high heat meat cooking, could also cause malignancies. Alcohol beverages containing ethanol are likewise capable of helping to create carcinomas in body.

Ultraviolet radiation is also part of the environmental causes. Light-skinned people are more prone to getting skin cancer from excessive sun exposure.

There are few materials that could cause malfunction of healthy cells, for example asbestos, which is dangerous in high dosage. Inhaling could form a lung carcinoma.

Other than that, some viruses could be origins, such as HBV (Hepatitis B virus), HCV (Hepatitis C virus), EBV (Epstein-Barr virus), HPV (Human papillomavirus) and HTLV (Human T-lymphotropic B virus).

One of the major risk factors that cannot be particularly avoided is genetic disposition through family history. One does not choose or change genome, it is information coded within body. If cancer has occurred in the family, the best way to prevent it are regular screening programs. It is not always just one factor responsible for causing cancer. Incidence of carcinoma can be formed with more than one, which leads to synergistic effects. Good example is alcohol, which reacts as a solvent for carcinogens located in tobacco products. [1] [3]

1.3 Cancer Incidence and Mortality Statistics

This disease is spreading worldwide and is therefore very important to monitor its global impact. The incidence, mortality and prevalence are the most important factors used in global cancer statistics. [7]

Incidence indicates the ratio of new cases of cancer, in given time period, to the total number of people in the population. This sum may be in the form of as an absolute number of cases per year or as a rate per 100,000 people per year. This number can be obtained from cancer registries.

Mortality is the result of the incidence and the fatality for a given type of cancer. In other words, it is the ratio of the number of deaths to a specific disease to the total

population. Its rate is therefore and indicator of the average risk of death caused by the disease. Mortality data are available via WHO (World Health Organization).

Prevalence is a term used for describing ratio of the number of people, who show a particular attribute of given disease, within specific time and location. In this case, it is not easy to determine the symptoms of the cancer, therefore people, who have been diagnosed with the disease, no matter if it happened long time ago or they had been cured, they all fall into this category. [7]

The next paragraphs will be dedicated to the statistics gathered by GLOBOCAN (Global Cancer Observatory) based on global estimates of cancer incidence and mortality, produced by the IARC (International Agency for Research on Cancer) for years 2002, 2012 and 2018. [7] [8] [9]

1.3.1 Global cancer statistics of year 2002

Throughout the year 2002, it has been estimated that there were 10.9 million new cases, 6.7 million deaths, and 24.6 million persons living with cancer, within 5 years since diagnosed. The most affected area on the body was lungs. The number of cases associated with lung cancer were approximately 1.35 millions and number of deaths were 1.18. Fatality of this type of disease gained value of 0.87. The following type, breast cancer, with number of new cases 1.15 millions and 411,000 deaths, had fatality only 0.35 due to good diagnostic procedures. Other widespread types were for example stomach and colorectal cancer. The difference in incidence and mortality of the most widespread types of cancer among men and women are shown on Fig. 1.2 [7]

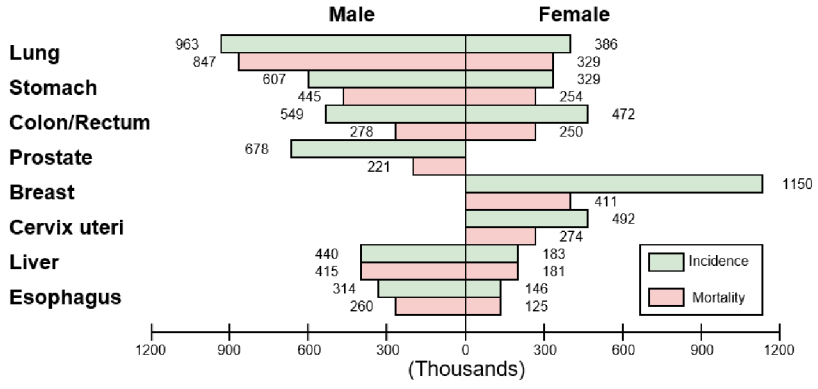


Fig. 1.2: Estimated numbers of incidence and mortality in 2002 according to sex and cancer site.

1.3.2 Global cancer statistics of year 2012

In 2012, the estimated number of new cases were 14.1 million and number of deaths 8.2 million. As seen on Fig. 1.3, the lung cancer is leading with 1.825 million of new cases worldwide, the fatality still standing at 0.87. Following, the breast cancer with number of new cases 1.677 million and the colon-rectum cancer with 1.36 million of new cases. Highest fatality is reached by liver cancer with value 0.95. The incidence related to women for esophageal cancer was not possible to obtain. [8]

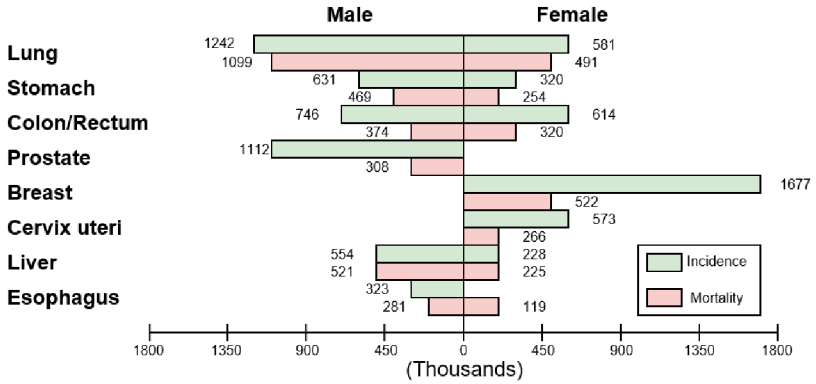


Fig. 1.3: Estimated numbers of incidence and mortality in 2012 according to sex and cancer site.

1.3.3 Global cancer statistics of year 2018

In comparison with previously stated years, year 2018 is leading in number of cases associated with cancer, which is 18.1 million. The estimated number of deaths affected by the disease is 9.56 million. The most reported cases were caused by lung cancer, approximately 2,094 million. Another widespread form is breast cancer. Colon-rectal cancer, whose number of new cases is 1.85 million, as seen on Fig. 1.4. The fatality of liver cancer has dropped, but it still has the highest value of 0.91. [9]

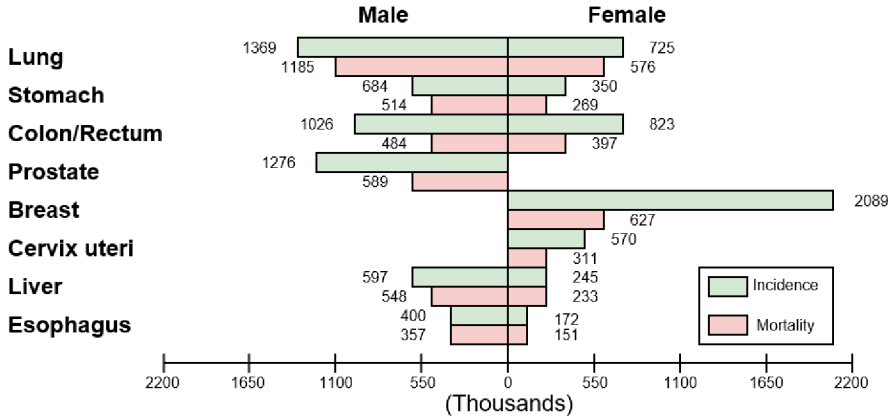


Fig. 1.4: Estimated numbers of incidence and mortality in 2018 according to sex and cancer site.

2 Methods

The analysis of the cells must go through various methods to ensure the required form for subsequent diagnosis. In this chapter, the light microscopy techniques used in cell imaging are described, where each type of microscope is being distinguished by its properties and therefore it is necessary to choose the right technique with respect to the research requirements. Later on, the image processing with a few exemplary methods, and ultimately a statistical method that will help interpret the results.

2.1 Light Microscopy

Light microscope is an optical device that can magnify the image of specimen. The principle is in visible light beam, which passes through the specimen and afterwards is bent in the lens system (Fig. 2.1). [10] [11]

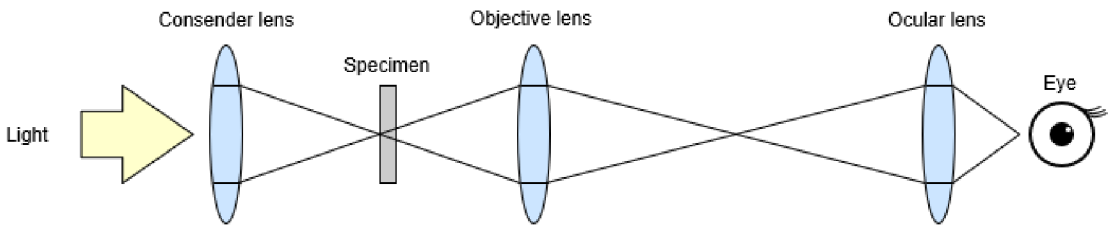


Fig. 2.1: Principle of Light Microscope.

The structure usually consists of ocular lens, tube, objective lens, stage, condenser lens, fine focus, coarse focus, iris diaphragm, light source and base. The objective lens, is the essential part of the microscope, responsible for resolving small details of the specimen and quality of the produced image. This structure is very difficult to construct, because it contains number of lenses of different shapes. Another important part is a condenser that focuses light beam into the place, where the specimen is. [10] [12]

2.1.1 Light Microscopy of Live Cells

There are many types of systems that can show live cells, but each works on a different principle. For these reasons, it need to be considered which system to use for imaging living cells. Emphasis is placed on sensitivity of detection, acquisition rate, and the viability of the specimen. For the system selection is also important

specimen properties, such as the thickness and characteristics of the process being studied, such as speed. [13] [14]

In the observation of cells, emphasis is placed mainly on their physiological properties, such as division, Brownian motion, diffusion, and a lot more. These properties occur in living cells, which means they cannot be fixed, freezed or dehydrated. Therefore, the light microscope is used that allows observation of live cells at high resolution and contrast, using special optical systems. [13]

Contrast Enhancement

Contrast is one of the primary image quality parameters, the source of which is the interaction of objects with illuminating radiation. One way to improve it, is staining. Using this technique, in addition to basic information about the physical morphology of objects, chemical composition is also obtained. There are many types of dyes and each of them has an affinity for specific cellular structures. Dyes that react differently to different kinds of cells are called differential colors, the basic example being the Gram Stain. Another types are fluorescent dyes which, after irradiation, can emit light in the form of fluorescence. An example is the dye DAPI (4',6-diamidino-2-phenylindole), which is used to visualize the cell in its natural environment. [13] [15]

Obviously, not all cells can absorb light. Cancer cells in tissues culture are fairly transparent but their components have specific refractive indexes causing a phase change of light, which can be detected by interference. These objects can be observed by two methods, namely direct phase shift detection, which is described in the phase-contrast microscopy, and oblique illumination. The principle of this method is to shade the condenser so that part of the light aims aside from the objective lens. Objects that have the property of changing the direction of the passing light, appear in the resulting image darker or lighter than the background. This aspect is caused by the part of a light that has either deviated even further from the objective lens or, on the contrary, has pointed towards it. [15]

2.1.2 Major types of light microscopy

Bright-field Microscopy

This type of microscopy is best known for its simplicity. Light passes through the specimen, which reduces its intensity under the sample absorption. The outcome is a dark object on a light background, therefore the name, bright-field microscopy. The used optics doesn't alter the color of the specimen, which is one of the advantages.

The disadvantage is that only a few samples can be observed without additional staining. [12]

Dark-field Microscopy

By using a special condenser lens and an opaque disc we can observe in dark field. The principle is in transition of light beam, while the disc is located in the middle of lens, resulting in removing the central part of light cone, leaving only rays located in the outer ring. The majority of beam is scattered when passing through the specimen. Afterwards it enters the objective lens and creates an image. The result is an bright image against the dark background. The advantage of this technique is significant increase in contrast compared to bright-field microscopy. It is suitable for viewing external features of cells, such as outlines, edges. Although, the observed cell must have the following attributes: it is transparent, doesn't contain added stains and absorbs little or no light. A great disadvantage is the tendency to distort the resulting image, causing inaccuracies. For example the artifacts may be caused by the insufficient specimen thickness. [10] [12] [13]

Phase-Contrast Microscopy

As seen on Fig. 2.2, a condenser annulus is positioned in the front focal plane of the condenser, causing parallel light wavefronts emanating from the ring through the specimen. Phase objects will divert some of the rays from the original direction. Phase plate located in the rear focal plane shifts phase of rays which have not changed their direction. The image of the object originates from the interference of the diffracted light (phase-shifted) and surround light. When the image compared to its background appears darker - *positive phase contrast*, when lighter - *negative phase contrast*. The main reason using this method is high contrast an image resolution. It is suitable for obtaining information from living cell in its natural habitat. A serious lack occurring during method is the presence of so-called halo effect which appears in materials with a high refractive index. It is a glowing interface between the object and the surrounding environment, as a result of a light refraction. The limitation may also be the size of the cell, whereas this method is not ideal for observing thick structures. [10] [11] [12] [15] [16]

Differential Interference Contrast Microscopy

Also called the Nomarski interference microscope, whose main element is the Nomarski prism. The light passes through a polarizing filter, falls on the Nomarski prism, where the light separates into two mutually polarized parts. After passing though the sample, they are shifted and before the impact on the image plate, they

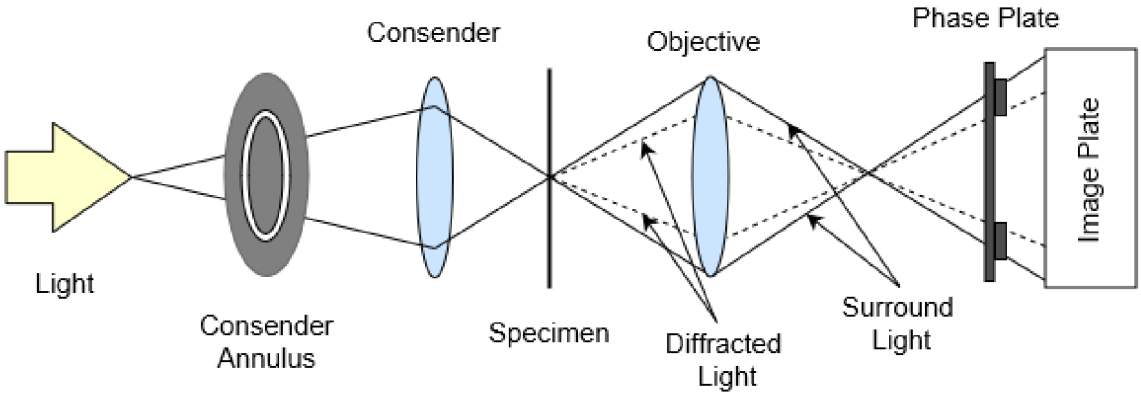


Fig. 2.2: Principle of Phase-Contrast Microscopy.

are recombined in the same way as they were separated from each other. It is mainly used in the visualization of living cells, due better image quality. The previously mentioned halo effect does not occur in this method, which allows better identification of external cellular structures. It can also observe extremely thin specimens and biological processes with additional staining. [12] [16] [17]

Fluorescence Microscopy

Frequently used for substances capable of emitting visible light after being illuminated with light of shorter wavelength, such as UV (Ultraviolet) irradiation. The light path begins with passing through excitation filter into dichroic mirror where the change of direction occurs. After light hits the specimen, electrons absorb the energy, causing them to excite into higher energy level. Being unstable, they return to basic state, accompanied by photon radiation, which is subject of detection. The construction is similar to a conventional optical microscope with adaptation to work with UV light. It is mainly used for observing cellular structures, thanks to its high sensitivity, which can detect even smaller molecules. The disadvantage is inability to examine thick structures and almost all cells need to be dyed. [12] [18]

Confocal Microscopy

The major difference between this type of microscopy and the other mentioned, is presence of the laser using spectrum of visible, UV or IR (Infrared) light. The light passes through pinhole aperture toward the objective, focusing rays into the focal plane of the specimen. Subsequently, they are reflected back into the objective lens. Light rays carrying information about the structure of the specimen are reflected by the dichroic mirror toward another pinhole aperture to the detector, where the information will be processed. Scattered light rays that did not reflect from the structure

are intercepted by the pinhole aperture, increasing quality of the image. The major advantage is the imaging speed with high quality. Another important feature is the ability to display image in three-dimensional resolution. With this method is possible to observe functions of complex structures such as neural networks. The disadvantage is that it also needs stained cells. [10][12] [16]

2.1.3 Quantitative Phase Imaging

This method uses the principle of interferometry (Fig. 2.3) to measure the optical field, using both amplitude and phase information. The output of this method is the map of the optical path length delays associated with the examined specimen. This image consists of information about the local thickness and the refractive index of the structures of the specimen. [19] [20] [21]

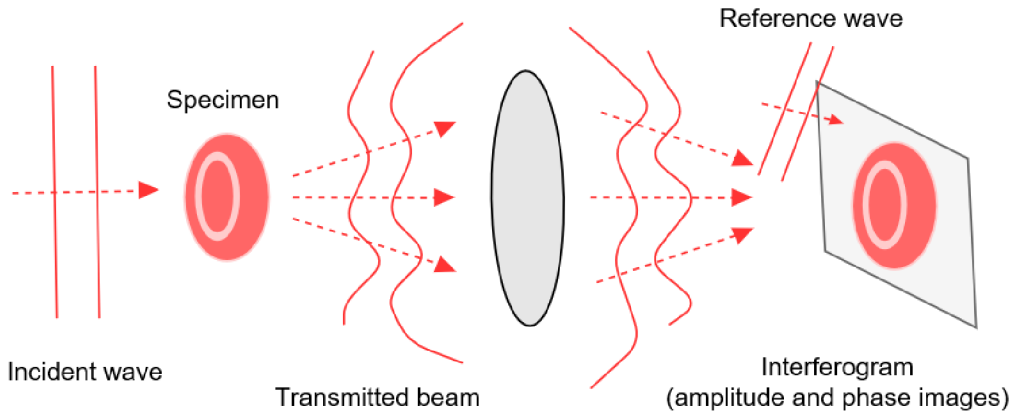


Fig. 2.3: Quantitative Phase Imaging operating on the principle of interferometry. [20]

Most biological samples, including cells, biomolecules and tissues, generally neither absorb nor significantly deplete the light beam. They are transparent, called phase objects. Meaning, the amplitude information does not provide a sufficient contrast for imaging, though they provide a phase delay, which serves as contrast for QPI (Quantitative Phase Imaging).

Unlike microscopic methods using additional dyes, QPI exhibits less phototoxicity and no photobleaching, it can even measure morphology and cell dynamics without added dyes. This is a huge advantage since the cells affected by staining may have altered its physiological functioning and cannot be inserted back into the patient's body. Another advantage is its sensitivity in nanoscale, for example, observing live neurons during electrical activity, is able due this ability. However, one of the most important reasons why to use this technique is its capability to measure

single-cell morphology such as volume or mass without damaging it, during any period of time. The observed cells may be in either adherent or flowing populations. It can also observe dry mass, which is the mass of a biological sample from which the water has been removed. [21]

From a medical point of view, this method has a great use already. Using QPI, cells do not need to be prepared or fixed by special procedures, also the labeling agents are not required. The main advantage is the already mentioned ability to monitor the morphology of the cell, according to which we can distinguish disease infected cells from healthy or even reveal the manifestations of onset diseases, whether infectious or noninfectious. QPI is also useful in studying the behavior of genetic diseases, for example, sickle cell anemia. By measuring the mean phase shift of cells, it is possible to determine whether the cell is dead, due to apoptosis or necrosis, or live, thus is moving. Another parameter this method can address is cell migration, for example by detecting the traction force created by fibroblasts during migration. [20] [21]

2.2 Cell Image Processing

Human examination of cells has been routinely used and contributes significantly to successful treatments. However it may have limited objectivity and accuracy in detecting small but systematic changes in dynamic cell behavior. That is the reason, why applications are used nowadays, to facilitate the extraction of quantitative information from cells that can determine the disease quickly and effectively, resulting in faster diagnosis. But before all that, it is necessary for the image of the examined cells to go through several processes. Cell image processing primarily consists of segmentation, tracking, feature extraction, and classification. [22] [23]

2.2.1 Segmentation

Segmentation is the most important and at the same time the most demanding process of whole cell image processing. It divides the image into several objects, regions, distinguishing them from the background. In medical applications, it can define areas representing, for example, bone structures, organs, cells. The level of difficulty of this process depends on the type of image being processed. [23] [24]

Segmentation techniques can be divided into the following groups:

- Parametric Image-based
- Region-based
- Edge-based [25]

Parametric Image-based methods operate on the basis of homogeneity of areas. The main procedure is thresholding. It is computationally simple and fast, dividing the gray-level image into binary by threshold, which can be determined by the histogram of the image. The use of this procedure is possible when the objects are not in proximal distance of each other and their gray-levels are visually distinct from the background. The parametric image-based methods naturally have their limitations, such as modification of the parameters depending on the illumination or the inhomogeneity of the image sensor. [25] [26]

Region-based methods are formed on the principle of continuity, which classify pixels based on their similarity to regions, areas characterized by a certain common characteristic. They use the concept of homogeneity, with more local application possible, resulting in greater flexibility in segment definitions. Homogeneity is an important feature that is used as the main segmentation criterion of the region growing method, which can be based on gray-levels, brightness, texture, etc. The region-growing method is one of the conceptually easiest methods of region-based segmentation and also probably one of the oldest segmentation methods. It is based on expanding areas depending on the fulfillment of a certain homogeneity criterion. Enlargement continues until the area is surrounded by a region with a different criterion. In this case, no further expansion is possible. The other method used, is a region split-and-merge, based on dividing regions, that are not homogeneous in the sense of the selected criterion, into smaller until all the chosen regions are uniform. Subsequently, the adjacent regions are merged. [25] [26] [27]

Edge-based methods are another way in image segmentation, using the principle of discontinuity. Their goal is to find the edges of the segments in the image. When the boundaries are correctly determined, the edges are formed by closed curves whose internal area is sought area of the examined segment. The methods also use homogeneity, but its requirements are low, meaning, the edge is detected in places where abrupt changes in image properties have occurred, which may be indicative of the potential boundaries of the region. The created curves should be closed, describing the circumference of examined the area. There are, of course, problems associated with edge-based methods, the most common of which are caused by noise or insufficient information in an image. These causes may result in segmentation in a way of edge being located where there is no border and vice versa no edge is located where a real border exist. One of the methods used, where image noise is present, is edge relaxation, using edges located between pixels, i.e. crack edges. [25] [26]

2.2.2 Tracking

The essential property of each object in the real world is its expansion in time and space. This feature is especially typical for living organisms that need time to grow, reproduce or even respond to stimuli. This is the reason why real-time cell analysis is used in medicine. However, one of the problems that may arise during the analysis of a large number of cells is object tracking. Imagine exploring hundreds to thousands of cells in a few hours film. It is almost impossible to manually trace individual cells in hundreds to thousands of frames. Therefore, sophisticated computer methods have been developed to fulfill this task. [28] [29]

Cell tracking methods can be generally sorted into two groups:

- Tracking by detection
- Tracking by contour evolution [30]

Tracking by detection is characterized by the principle of segmenting all objects in all frames in a video which associate objects among frames based on similar properties. The most commonly used method is the *nearest neighbor (NN)* algorithm that is simple, efficient and in many cases effective. The principle is to link each segmented cell in a given frame to the nearest cell in the next frame. The term nearest may refer to spatial distance but also to differences in properties such as intensity or volume. This method works seamlessly if the cells are well separated in at least one of the dimensions of the feature space. Naturally, limitations are associated with this method, meaning high memory requirement and high computation complexity, which led to creating techniques capable of overcoming mentioned limitations. NN techniques are broadly classified into structure less methods (Reduced NN, Condensed NN), overcoming memory limitation and structure based methods (Tunable NN, Nearest Feature Line), reduction in computational complexity. [28] [31]

Another method we can include in tracking by detection is graph-based optimization, which perceives the image as a graph. Each pixel represents a node and edges to reflect changes in image properties, such as intensity or color. [32]

Tracking by contour evolution methods begin by segmenting the cells in the first frame and evolve their contours in the next frame. The advantage of this method is simultaneous segmentation and tracing of objects, assuming spatio-temporal overlap of corresponding cellular regions. The method accurately represents the silhouette of the object and, during frames, directly evolves the contour of the object by minimizing energy using direct minimization techniques, such as gradient descent and the region feature. Simply put, this technique uses the object contours of the previous frame as initialization in the current frame (Fig. 2.4). Another advantage is that it handles the deformation of the cells among the frames well, but if a new

object appears or there is a large displacement between the frames, reinitialization is needed. [30] [33] [34]

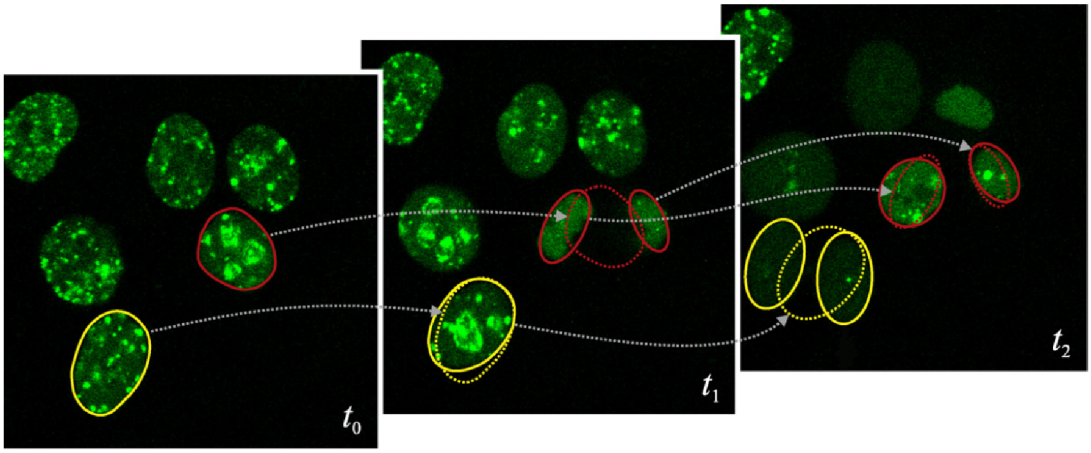


Fig. 2.4: In each frame in the time-lapsed image we see the final contour of a cell (solid line) obtained by minimization of an energy functional. The initial contour of a cell (dotted line) is used as initialization in the following frame. In the frames, mitosis can also be seen, which is detected by monitoring the shape of the contour function during energy minimization. [29]

2.2.3 Feature Extraction

Features are specific properties of the objects on which we can classify them. Two of the most informative features in medical image interpretations are shape and texture that can be quantified by many automated techniques. This allows feature extraction to be divided into two groups:

- Shape quantification techniques
- Texture quantification techniques

In the case of time-lapse image, dynamic features can also be used.

Shape quantification techniques

Shape quantification techniques can be applied to segmented images in three different ways. The first method measures the shape using geometric and statistical computations on all pixels within a segmented area, such as compactness and spatial moments. The second way to measure the shape is by means of boundary pixels to which geometric, statistical and spectral computations are applied. These include radial distance measures, chain codes and Fourier descriptors. The way to obtain

the shape of a cell that is elongated or branched is by media lines in which the essential shape information of the cell is stored. This information can be obtained using thinning algorithms. [35]

Compactness

Compactness C shows the similarity of the object to its smoothest shape, that is, a circle. It uses the perimeter P and the area A of the segmented region.

$$C = p^2/A \tag{2.1}$$

Compactness has no unit and its value increases with increasing shape complexity of the object. For the perfect circle it has a minimum value (4π). The compactness has been used, for example, for quantification of calcifications and breast tumors.

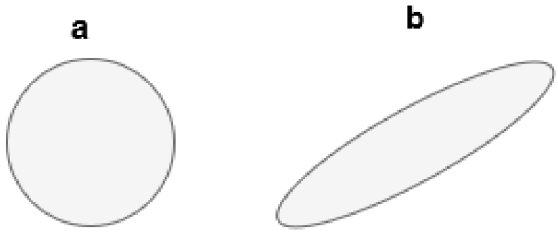


Fig. 2.5: Two binary objects whose compactness is **a)** 4π and **b)** 26,7. [35]

Circularity

Circularity is one of the most used shape factors. This feature is similar to compactness, but its maximum value is 1, indicating a perfect circle. The smaller the value, the greater the irregularity in shape.

$$f = \frac{4\pi A}{p^2} \tag{2.2}$$

where f is the shape factor, A means area and p is the perimeter. [36]

Elongation

Elongation is another feature for interpreting the shape of an object. It describes the difference from circular symmetry after plastic deformation. It is also referred to as the aspect ratio, where the length and width of the object is put in proportion. If the elongation is 1, in other words, the length and width of the object are the

same, there is no deformation in the object. The elongation can also be interpreted in such a way that if the value is less than 1, the object has a width greater than length, and if the elongation value is greater than 1, the object length is greater than width.

$$\text{elongation} = \frac{a}{b} \tag{2.3}$$

where a means axis A, length of object and b means axis B, width of object. [36]

Skew

The term skew describes the measure of asymmetry, which indicates how distribution of values differ from the normal distribution. In other words, it indicates whether the distribution of values comparing to a normal distribution is symmetrical or asymmetrical. In image processing, skew is obtained from a histogram of grayscale intensities. This histogram contains individual pixel intensities of the object. The value of skew can differ. If the value is 0, it means that the histogram layout is symmetrical. If the value is less than 0, it is a negative skew, which means that the mean value is less than the median and thus the "tail" of the distribution is on the left. If the skew value is greater than 0, it is a positive skew, which means that the mean value is greater than the median and hence the "tail" of the distribution is on the right side.

$$\text{skew} = \frac{1}{\sqrt{\frac{1}{A}\iint_A \left(Z^2(x,y)\right) dxdy}^3} \left(\frac{1}{A} \iint_A \left(Z^3(x,y) \right) dxdy \right) \tag{2.4}$$

where A means area of the object and *Z(x,y)* means a deviation from the mean value. [37][38]

Spatial moments

In addition to statistical analysis, these moments can also be used to represent spatial distribution of values in a 2D function. Moments of a M by N image *f(i,j)* are given by

$$m_{pq} = \sum_{i=0}^{M-1} \sum_{i=0}^{N-1} i^p j^q f(i,j) \qquad p,q = 0,1,2,3,\ldots \tag{2.5}$$

where p+q is the moment order of two dimensions.

For example, *translation invariance* can be achieved by using central moments

$$\mu_{pq} = \sum_{i=0}^{M-1} \sum_{i=0}^{N-1} (i-\bar{i})^p (j-\bar{j})^q f(i-\bar{i}, j-\bar{j}) \qquad p,q = 0,1,2,3,\ldots \tag{2.6}$$

where

$$\bar{i} = m_{10}/m_{00} \qquad \bar{j} = m_{01}/m_{00} \tag{2.7}$$

It is also possible to obtain the *orientation* of the object as the direction where it is most elongated. This information may be obtained with the angle θ

$$\theta = \frac{1}{2} \tan^{-1} \frac{2\mu_{11}}{\mu_{20} - \mu_{02}} \tag{2.8}$$

To obtain *eccentricity* ε, which is defined as the ratio of the main axis and the perpendicular axis. It ranges from 0 to 1, with 0 indicating that the object is a circle and 1, a straight line

$$\varepsilon = \frac{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^4}{(\mu_{20} - \mu_{02})^2} \tag{2.9}$$

The spatial moments can be used for quantification of breast tumors or mammographic calcifications. For objects smaller than 15 pixels or when parts are relatively small, the efficiency of spatial moments deteriorates. [35]

Radial distance measures

Another way to determine the shape of an object is by using a radial distance $d(n)$ that is measured from the centroid (x_c, y_c), in the given area, to each pixel $(x(n), y(n))$ on the boundary,

$$d(n) = \sqrt{[x(n) - x_c]^2 + [y(n) - y_c]^2} \qquad n = 0, 1, \ldots, N-1 \tag{2.10}$$

where N means number of pixels. [35]

Chain codes

The shape of the object can be determined by a chain code technique that analyses each point at the boundary in sequence, assigning each of them the digital code to the transition. Prior to using this technique, it is necessary to adjust the boundary by reducing the spatial resolution with a new x-y grid. The reason for this preprocessing is to limit the minor deviations that are considered to be part of the object, due to noise or segmentation pitfalls. The most common types are 4-connectivity, taking into account the 4 closest neighbors, indicating their numbers from 0 to 3, or 8-connectivity (Fig. 2.6), taking into account all adjacent points, indicating their numbers from 0 to 7. Chain codes has been used in numerous analysis, such as neural tissue analysis or analysis of single and overlapping lymphocytes. [35]

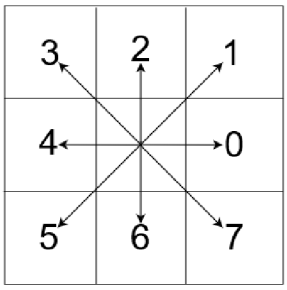


Fig. 2.6: Transition label definitions using 8-connectivity chain code technique. [35]

Fourier descriptors

The emphasis of this method is placed on the contour c, where each pixel on the contour of an object can be represented by a complex number whose real and imaginary parts are the x and y coordinates of the selected pixel. This can be obtained by tracing around it in the selected direction, starting from a selected pixel,

$$c(n) = x(n) + jy(n) \qquad n = 0, 1, 2, \ldots, N-1 \tag{2.11}$$

where n is index of the pixel, N is number of pixels and j means imaginary unit number.

After applying a discrete Fourier transform (DFT) to the contour sequence, the spectral coefficients can be obtained, which contain all the shape information about the object.

$$d(u) = \frac{1}{N} \sum_{n=0}^{N-1} c(n) e^{-j2\pi nu/N} \qquad u = 0, 1, \ldots, N-1 \tag{2.12}$$

The essential shape information is typically contained in low-ordered coefficients. For example, the first coefficient *d(0)* contains information about centroid of the contour. However, it should be noted, that all the coefficients dependent on the first pixel chosen as the starting point. Fourier descriptors have been used in the medical field to recognize and represent different parts of the human body, such as the recognition of human corneal endothelial cells or the represent the shape of prostate in magnetic resonance images. [35]

Texture quantification techniques

When examining the medical images, emphasis is often placed on the structure of the examined tissue, which can be described by properties such as smoothness, grain, homogeneity. The structure, unlike the shape, describes the entire selected region. One of the methods by which the texture can be estimated is statistical moments.[35]

Statistical moments

Statistical moments can be used as a texture metrics. They are based on histograms, where the histogram *h(k)* indicates the number of K gray levels ranging from 0 to K-1. Objects that have a smooth texture have pixels value close to each other, on the other hand, rough texture objects have wide variability in pixel values. For example, the second moment m_2 indicates the variance, which is often used indicator, due to its good correlation with the visual roughness perception and it has relatively low computational complexity. The third moment m_3 means a skewness, which reflects asymmetry, and the fourth moment m_4, the information about the kurtosis, that is, the uniformity. The nth moment for a given histogram is estimated by:

$$m_n = \frac{1}{K} \sum_{k=0}^{K-1} (k - \mu)^2 h(k) \tag{2.13}$$

where μ is mean value of gray levels.

Statistical moments, as previously mentioned, are based on the histograms, meaning they contain no spatial information. However, they have been used for applications such as texture quantification in ultrasonic images of liver tissue. [35]

Dynamic features

In the field of biomedical research, the dynamic features of cells are determined. Cellular dynamic morphology is cellular deformation over a period of time. It can be calculated by *a cellular contour*, whereas its spectrum is composed to characterize spatiotemporal deformation. The contour is represented by radial distance and is used primarily to characterize cell morphology. The contour function C is calculated based on the centroid's position change (x_c, y_c),

$$C(k,t) = \sqrt{(x_{k,t} - x_c)^2 + (y_{k,t} - y_c)^2} \quad k \in [0, 360] \quad t = 1, 2, \ldots, T_n \tag{2.14}$$

where $(x_{k,t}, y_{k,t})$ mean coordinates of radial distance sequence, k means an angle, t is a frame index and T_n denotes the length of image sequence n. [39]

Protrusion

The protrusion is part of the object in the newly allocated space by the object in the subsequent frame compared to the previous one. It is computed using masks, one for the subsequent frame and the other for the previous one. An important step is to binarize and afterwards subtract these masks, while only the positive values describe the protrusion parameter.

$$\text{protrusion space} = (Mask_{(i+1)} - Mask_{(i)}) > 0 \tag{2.15}$$

where $Mask_{(i+1)}$ means mask of the subsequent frame and $Mask_{(i)}$ means mask of the previous frame.

Area can be computed from the number of pixels located in the newly allocated space.

The *protrusion max value* is the largest protrusion value between pixels in the newly allocated space.

2.2.4 Classification

Objects obtained from the segmentation can be divided into subgroups called classes which properties are, from a classification point of view, similar. Object recognition is based on assignment to certain classes. In charge of this process is a classifier, that assigns objects to appropriate classes based on their properties, called a pattern. This means that the classifier assigns the objects from their patterns, not from objects itself.[26]

Pattern recognition techniques are broadly categorized into two groups:

- Classification methods based on supervised learning
- Classification methods using unsupervised techniques

The supervised methods can be additionally divided into:

- Parametric classifiers
- Nonparametric classifiers [22]

Supervised parametric classification uses labeled pattern samples to train the classifier in order to estimate the statistical parameters of each class of the pattern. The term labeled pattern samples means patterns which membership in the class is previously assigned. Frequently used methods of this type are the minimum distance classifier or maximum likelihood classifier.

Supervised nonparametric classification, unlike the previously mentioned, does not use parameters to assign objects into classes. One of the most used techniques is the K-nearest neighbor.

Unsupervised classification separates the entire set of data based on a similar criterion, resulting in the apportion of objects into clusters. [22]

There are many classifiers, but in the following paragraphs there will be described only a couple of those, which are the most notable, such as Artificial Neural Network (ANN) and Support Vector Machines (SVM).

Artificial Neural Network

Artificial neural network, in short neural network, is structure composed of many nonlinear computational elements, nodes, arranged in a pattern resembling biological neural network. The desired behavior of the ANN is derived from a process known as a learning period that can be supervised (to each input of the vector X is assigned a known output vector Y) such as back-propagation network or unsupervised (based on automatic recognition) such as Kohonen network. The computational elements themselves are usually interconnected by weights which are adapted to the given application. These weights can adapt over time, based on current results, which improves the performance of the entire neural network. Damage to the computational element will not significantly affect the overall performance of the network. ANNs are often used in speech and image recognition. Classifier based on neural networks is, for example, back-propagation which will be described in more detail. [22] [25]

Back-propagation network, also called a multilayer feedforward network, is composed of multiple layers of the neural network, see (Fig. 2.7). The information spreads from the input layer to the others up to the output layer, with no response delay. Layers between input and output are called hidden layers. The weights of each node are adjusted to provide an approximation of the required mapping, typically by supervised learning using error backpropagation. The error spreads with every learning step back, from output layer to each lower layer. This process makes it possible to determine an individual error for each neuron at each step. The error can be calculated,

$$\delta = y_d - y \tag{2.16}$$

where y_d is the desired output and y is the actual output. When a local error is known, the weights can be adjusted using the delta rule,

$$w_i(t+1) = w_i(t) + \mu\delta x_i \tag{2.17}$$

where $w_i(t+1)$ is newly calculated weight, $w_i(t)$ is actual weight, μ is learning rate and x_i is actual input. After hundreds to thousands of learning steps in this process, the resulting mapping performed by the network usually corresponds to the desired output. In order to reach a large number of learning steps with a training set of adequate size, learning epochs are used. [22] [25]

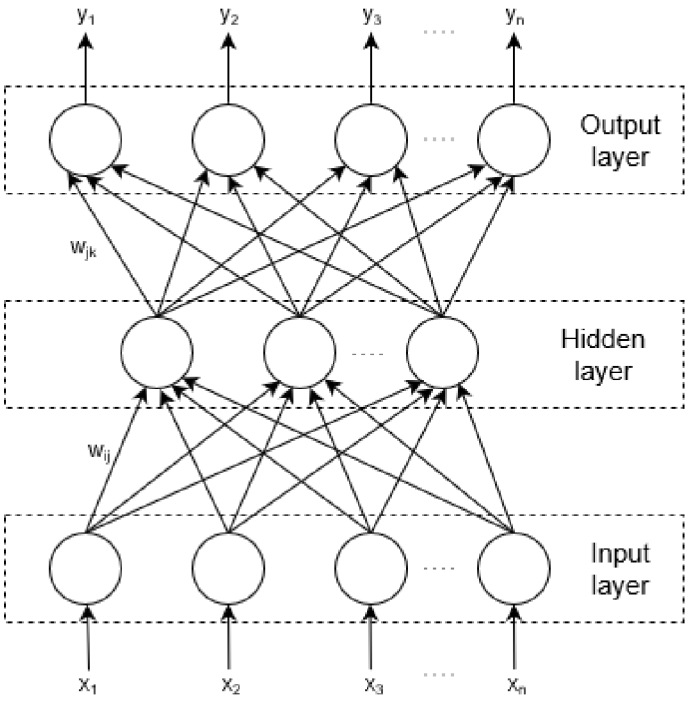


Fig. 2.7: Architecture of a three-layer backpropagation network. [22]

Support Vector Machines

Support vector machines is a supervised machine learning technique that divides data into two groups based on maximizing the width of empty area, called margin, between them, known as the maximum margin hyperplane. The width of the margin is defined as the distance between the hyperplane and the closest training patterns, called support vectors (Fig. 2.8). SVM works well in a number of biological analyzes, particularly as a classifier for cellular applications. The original classifier was developed for the linear separation of two groups, later, after various modifications, it was possible to use nonlinear separable classes or non-separable classes. [26]

Let assume the simplest case and that is *the two-class linearly separable case* with training data ($\vec{x}$, $\vec{y}$), where $\vec{x}$ is the feature vector and $\vec{y}$ is class identifier vector, where, for simplification, $\vec{x} \in$ [0,1] and $\vec{y} \in$ {-1,1}.

Equation of the hyperplane is:

$$\vec{x} \cdot \vec{w} + b = 0 \tag{2.18}$$

where $\vec{w}$ is a normal vector of a hyperplane.

To maximize margins, it is necessary to define two parallel hyperplanes passing through support vectors, with no training pattern between them

$$\vec{x} \cdot \vec{w} + b = 1 \tag{2.19}$$

$$\vec{x} \cdot \vec{w} + b = -1 \tag{2.20}$$

To ensure that no training patterns are present between the two hyperplanes, all patterns $\vec{x}_i$ must meet the following inequality

$$\vec{y}(\vec{w} \cdot \vec{x}_i + b) \geq 1 \tag{2.21}$$

Therefore, the distance between these two hyperplanes acquires value of $(2/||\vec{w}||)$. [26]

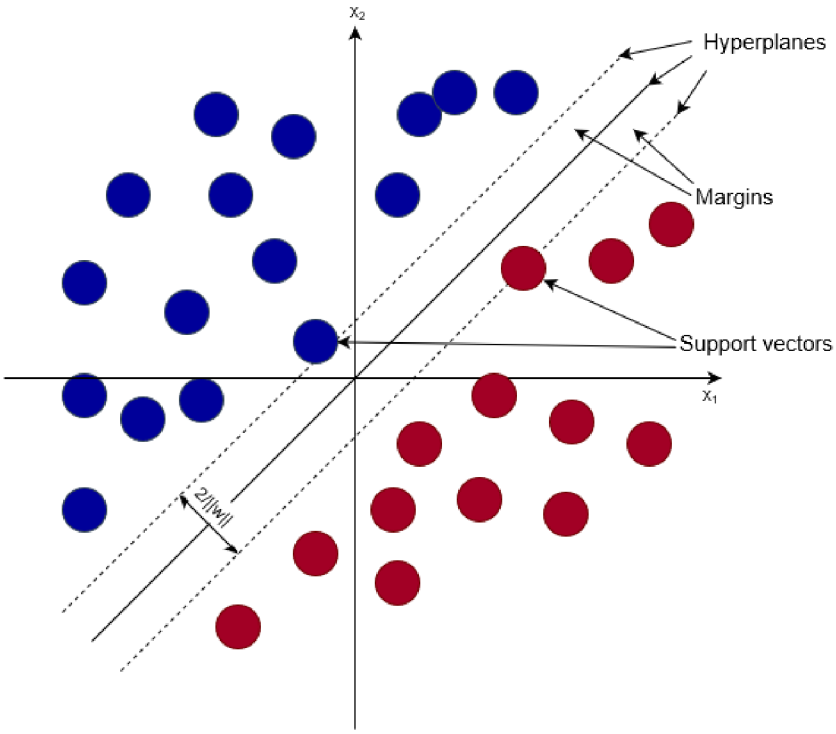


Fig. 2.8: Two-class classification using support vector machines.

2.3 Principal Component Analysis (PCA)

Principal component analysis is a mathematical algorithm whose main advantage is reducing data dimensionality. It is one of the oldest and most widely used methods of multivariate analysis. It also has the advantage of reducing the number of characters (data), while avoiding large loss of information content. Dimension reduction is created by recognizing the direction in which the largest variability of the measured data is located. This direction is called the *principal component*. The first principal component contains the largest data variability, the second also contains data variability, but only those which was no longer contained in the first. The third principal component contains such a variance that has not been included in the foregoing and continues in this way until no principal component can be formed. These components are sorted by importance, that is, by decreasing variation from the largest to the smallest. This method identifies new variables that are a combination of input parameters. [40] [41]

The PCA calculation consists of several steps. If we want to fully understand it, it is necessary to define the basic elements found in this method beforehand.

The first is *covariance*. This term describes dependencies between two dimensions (variables). If we apply a covariance to one and the same variable, we get a variance. If we measure two variables *(x, y)*, we get a covariance between x and y. To calculate the covariance, this formula can be used.

$$cov(X,Y) = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \tag{2.22}$$

where X_i and Y_i are input data, $\bar{X}$ and $\bar{Y}$ are its mean values and n stands for number of data.

It is also important to add that the input data must be the same length, meaning they have to have equal number of values. For the correct interpretation of the result we are not so interested in the value itself, rather its sign. If the result is positive, the value of both dimensions (variables) increases. If the result is negative, then as one dimension (variable) increases, the other decreases. It is possible that the covariance value may be 0, so the dimensions (variables) are independent. After applying several variables, for example three *(x, y, z)*, we get a covariance between x and y, x and z and y and z.The more dimension we want to measure, the more covariance results we will have. For the n-dimensional dataset we get numerous different covariance values. Therefore, a so-called *covariance matrix* is introduced for a multi-dimensional calculation. The principle is to calculate the covariances of the n-dimensional dataset and then insert them into the matrix. For a better understanding, the covariance matrix is depicted using the x, y and z dimensions.

As it is shown the matrix is symmetrical about the main diagonal.

$$C(x,y,z) = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

Other terms that need to be defined are *eigenvalues* and *eigenvectors*.

Eigenvectors are non-zero vectors whose direction does not change after transformation. An important feature of these vectors is that they are orthogonal, i.e. rectangular. The coefficients that multiply the size of eigenvector are called eigenvalues. The mentioned eigenvector and eigenvalues can only be obtained from square matrices. It is important to note that are always in pairs. In the case of (N x N) matrix we get N eigenvectors and N eigenvalues.

After explaining the main terms used in the PCA method, it is best to go to an explanation of how this method actually works. The first step is to determine the data to work with. The input should be a matrix where the dimensions (variables) are inserted in the columns. The dataset needs to be modified in a way that their values are standardized, ie their mean is 0. Subsequently, the covariance matrix from which we get our eigenvectors and eigenvalues is calculated. With these two variables (eigenvectors and eigenvalues), we can detect the principal components so that the eigenvector with the highest eigenvalue is the principal component of the dataset. After obtaining eigenvectors and eigenvalues from the covariance matrix, the next step is to sort these vectors by eigenvalues from highest to lowest. This indicates how significant the individual components are. The final step of the PCA method is to create new dataset. This step is created by multiplying the transposed matrix of the dimensions with the transposed matrix that contains the eigenvectors. The result is a new dataset that combines the previous input dimensions. [42]

2.4 Statistical Methods

Medical science and its research are based on knowledge that consists of hypotheses from known and unknown areas. It is by using statistical methods that analyze the collected health-related data to draw valid conclusions regarding hypotheses. Statistical methods can be generally sorted into two groups:

- Descriptive statistics
- Inferential statistics

Descriptive statistics deals with simplifying data for better understanding and to interpret the pattern found in analyzed samples. Since medical research and clinical observations continually examine collected data, it is important to decide whether

the pattern and characteristic in data sets is unusual or happened at random. This step is crucial in the subsequent application of the findings into clinical application.

Inferential statistics helps with these decisions, including hypothesis testing.

As the emphasis in medical field is placed on variability and uncertainty, statistical methods have been developed to help in taking the decision regarding to the data collected with variability. Thus, statistical methods are necessary in order to solve this problem. One of the techniques for data analysis with variability is method called the analysis of variability (ANOVA), which will be described in more detail. [43]

2.4.1 Analysis of variance (ANOVA)

Analysis of variability is part of the statistical parametric methods. It is used to analyze and compare data from experiments, while best being used when comparing more than two populations or samples. It is subject to considerable complexity and subtlety, with many variations of different type, each of which is applied in a specific experimental context. ANOVA has three types depending on number of the factors, independent variables:

- One way analysis - comparing groups based on one factor
- Two way analysis - comparing groups based on two factors
- K-way analysis - comparing groups based on k number of factors

It is also important to mention the concept of hypothesis. In the case of ANOVA, a null hypothesis (H_0) means that all the mean values of the k number of groups are the same and the alternative hypothesis (H_1) means for at least two groups to differ.

$$H_0 = \mu_1 = \mu_2 = \ldots = \mu_k \tag{2.23}$$

$$H_1 = \mu_1 \neq \mu_2 \tag{2.24}$$

But before this method can be used on data, the following assumptions must be met:

- Homogeneity of variance
- Independent observations
- Normality

Homogeneity of variance, or homoscedasticity, should be similar within each groups. Its calculation is needed when the sample sizes are unequal. The two most used tests of homogeneity are by Levene and Brown & Forsthy.

Independent observations should be independent or a sample should be randomly selected. No significant or any pattern should occur in the selection of the sample.

This problem can be avoided by random sampling, random assignment to groups, and experimental control of extraneous variables.

Normality means that each group, and the subgroup should be normally distributed. If a small sample size is analyzed, normality should be calculated using the Shapiro-Wilk test, with the sample size N<50, or D'Agnostino's modification, with the sample size N>50. [44]

ANOVA is based on general linear models which are:

- Fixed effects model
- Random effects model
- Mixed effects model [45]

Fixed effects model makes conclusions valid only for the analyzed data. These conclusions are specific for the given population and treatment study. The levels of the independent variables are fixed depending on the design of the experiment. For a better idea, if four different drugs are applied to four treatments, conclusions can only be drawn for these specific drug doses.

Random effects model draws conclusions about independent variables that are not used depending on the experiment, which means that it makes inferences about a population's random variation. For example, in the previously mentioned case about four different drugs, the random model would conclude on the continuum of drug doses which is not used in the study.

Mixed effects model consists on both fixed and random effects. [44] [45]

Simple ANOVA procedure

The basic outcome of the analysis of variance is the ANOVA Table (Tab. 2.1), consisting of the following terms that will be described.

Sums of Squares (SS) measure the difference of the mean value of a given group to the overall mean value. In the case of ANOVA, the computed sum of squares can be between the groups SS_B, within the groups, also called the error SS_E, which is calculated by the sum of the squared differences between the individual observations and their group means. Also, the total sum of squares SS_T is calculated, which is estimated by the sum of the squared differences between the individual observations and the overall mean value.

$$SS_B = \sum n_j(\bar{X}_j - \bar{X})^2 \tag{2.25}$$

where n_j means sample size per group, $\bar{X}_j$ is sample mean of the jth group and $\bar{X}$ is overall sample mean.

$$SS_E = \sum \sum (X - \bar{X}_j)^2 \tag{2.26}$$

where X means individual observation.

$$SS_T = \sum\sum(X - \bar{X})^2 \tag{2.27}$$

Degrees of Freedom (df) indicate the number of values that in the final computation of statistics can freely varies. In the case of *df* between groups, it is calculated as the number of groups k - 1, in the case of within groups, the *df* as a difference of the total number of observations N and the number of groups k. Total degrees of freedom are calculated as the number of observations N - 1. For better understanding see Tab. 2.1.

Mean Squares (MS) is the term used to estimate the variance between the groups, which is calculated as the ratio of the sum of squares and the its corresponding value of degrees of freedom. In ANOVA, two types of *MS* are used: the mean squares between the groups MS_B and the mean squares within the groups MS_E.

$$MS_B = \frac{SS_B}{k-1} \tag{2.28}$$

where k means the number of groups.

$$MS_E = \frac{SS_E}{N-k} \tag{2.29}$$

where N means total sample size or total observations.

F–test is used if we want to know if the estimates from two sources of variation are different. The F-value is calculated of the mean squares between the groups divided by the value of the mean squares within the groups.

$$F = \frac{MS_B}{MS_E} \tag{2.30}$$

P-value is used to determine if any of the differences between the means are statistically significant. This value indicates the probability that if the null hypothesis is true, the statistical summary will be greater or equal to the actual observed result. A significance level, also referred to as α, is usually 0.05 in most statistical surveys. This value gives a 5% risk that the difference in data will occur even though there should be no difference. If p-value is less or equal than the significance level, the mean values are not the same, the null hypothesis is rejected and the results are statistically significant. If p-value is less than the significance level, the null hypothesis cannot be rejected and therefore the results are not statistically significant. [46] [47]

Tab. 2.1: ANOVA table with p-value

Source	SS	dF	MS	F	p-value
Between groups	SS_B	k-1	MS_B	F	p-value
Within groups	SS_E	N-k	MS_E		
Total	SS_T	N-1			

3 Results

In this part of the work, we will discuss the results of research of new potential biomarkers, that could be used for personal treatment of cancer. Cellular features which are described are both static and dynamic parameters.

The examined cells in this work come from line A549, which are adenocarcinomic alveolar basal epithelial cells, provided by MVDr. Pavlína Turánek Knötigová, Ph.D. from the Department of Pharmacology and Immunotherapy from the Veterinary Research Institute, v. V. I. in Brno. The drug that is applied to the cells is called GDC0941, an inhibitor of phosphatidylinositol-4,5-bisphosphate 3-kinase (PI3K). This enzyme is involved in multiple cellular features, including its motility and growth. The drug was provided by Julian Downward of the Francis Crick Institute, London, UK.

After recording the groups of drug-free cells and drug cells with 5 minute intervals, the images were subjected to a segmentation and filtering algorithms to allow consecutive analysis of the cells. Each cell was individually analyzed, meaning that each cell from each frame and each movie was detected by tracking algorithm. After determining which cell in the frames is the same one, its individual parameters have been obtained between frames and movies. This procedure was applied both to a control group of cells and to a group of cells containing drug GDC, in order to determine its accuracy. In order to determine whether the difference is significant, ANOVA method was used. The p-values in this work are represented by a graphical symbol (*), where (*) represents $p < 0.05$, (**) $p < 0.01$, and (***) $p < 0.001$.

At the end of each biomarker its classification by SVM algorithm is described. From each group, including the control and with applied GDC, a few cells were used As a training data to train the SVM classifier to recognize the individual groups between them, were for gdc group used 28 cells and for control group 33 cells. As a testing data were used 35 cells from gdc group and 36 cells from control group. However, the results may be distorted because it is possible that not all cells in the control group are infected with cancer.

The accuracy (ACC) of model classification in each biomarker is expressed using the terms *TP (True Positive)*, which indicates the number of positive detections, in this case, where the model detected a cell belonging to group with GDC, *TN (True Negative)*, the number of negative detections, the model detected the cell belonging to control group, *FP (False Positive)* , the number of positive detections, which are false, meaning the model classified the cell from control group into group with GDC and *FN (False Negative)*, where model classified the cell from group with GCD into

the control one. To get a percentage value, the equation will be used

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} * 100 \tag{3.1}$$

3.1 Analysis of speed

Speed is an important feature of cells. Motile cancer cells can move to different parts of the body through blood and lymphatic pathways and form metastases which is the most dangerous aspect of the disease. It is therefore an important aspect of chemotherapy to reduce the speed of cells.

After applying the algorithm designed to compare speed of the groups of cells, a p-value of 0.0000247218 was obtained which corresponds to variations between the GDC group and the control one. Thanks to this value, it is possible to conclude, that the results are highly significant.

The Fig.3.1 and the Tab. 3.1 show a significant decrease in mean and median value of cell motility after application of the GDC drug.

Tab. 3.1: Table with the results of comparing the speed of cell groups with the drug (gdc) and without the drug (cnt).

Group	Movies	Cells	Mean	SEM	Min	Max	Median	ANOVA
cnt	4	69	1.570	0.047	1.474	1.66	1.574	
gdc	6	64	0.386	0.074	0.191	0.674	0.375	$***$

The classification of this biomarker was up to 84% effective (3.2), indicating that the speed between cancer cells and cells with GDC drug is rapidly differentiated.

Tab. 3.2: Classification table of speed.

Number of cells	TP	TN	FP	FN	Accuracy (%)
68	32	26	7	3	85.3

3.2 Analysis of protrusion area

This biomarker refers to the cell movement. The more active the cell is, the larger the area, whereas the stagnant cell has a smaller area. The reason to indicate this biomarker as potencial, is because of previously mentioned fact, that the cancer cells are more active in motility.

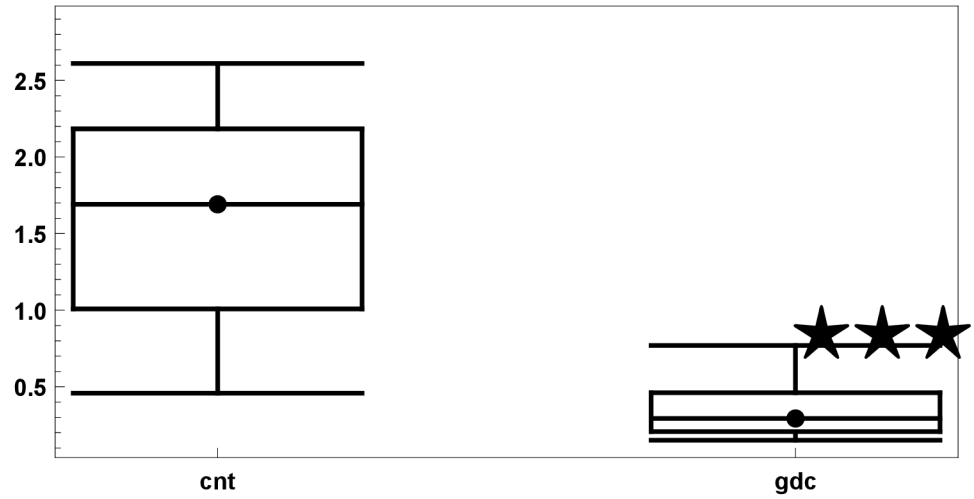


Fig. 3.1: Box and whisker plot of speed displaying significance between control group and group with gdc, where dot represents median, box represents 50% of data and whiskers 80%.

After applying the algorithm designed to calculate the protrusion area on both cell groups, a p-value of 0.00145804 was found, indicating variability between them.This value represents, that the result is significant.

As you can see in the Tab.3.3, the values of mean and median of gdc group are almost 4 times smaller comparing to those of cnt group. The Fig 3.2 describes the cnt and gdc group layouts, where it is clear that the values of gdc group are smaller.

Tab. 3.3: Table with the results of comparing the protrusion area of cell groups with the drug (gdc) and without the drug (cnt).

Group	Movies	Cells	Mean	SEM	Min	Max	Median	ANOVA
cnt	4	69	486.84	47.4	3.662	2481.7	365.42	
gdc	6	64	148.82	22.51	6.871	850.04	94.343	**

Values obtained after SVM classification of cells can be seen in the Tab3.4. The accuracy of recognizing individual cell membership in groups was 78%.

Tab. 3.4: Classification table of protrusion area.

Number of cells	TP	TN	FP	FN	Accuracy (%)
68	21	32	1	14	78

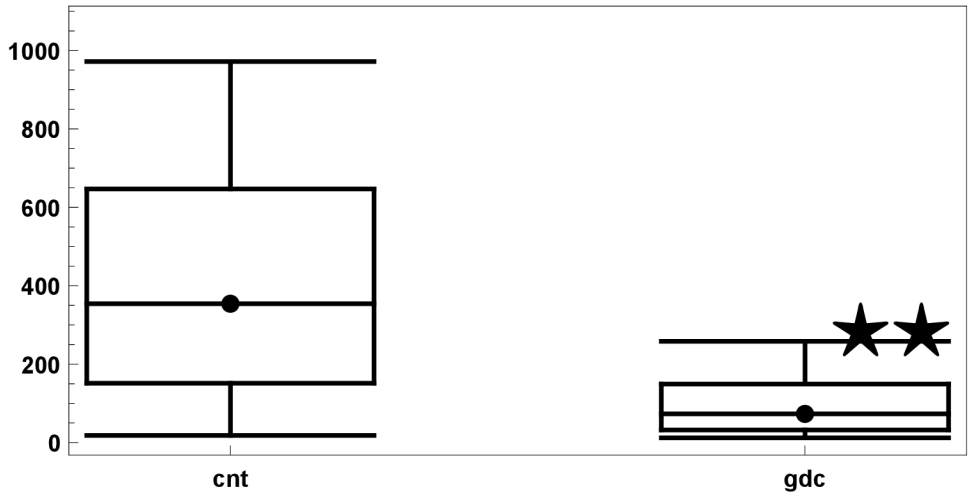


Fig. 3.2: Box and whisker plot of protrusion area displaying significance between control group and group with gdc, where dot represents median, box represents 50% of data and whiskers 80%.

3.3 Analysis of protrusion max mass

As the aforementioned biomarker, protrusion area, this one is also related to cell membrane variability, but instead of its area it shows the value of the maximal mass present in the cell. This parameter is also potentially useful because cancer cells are constantly multiplying, indicating a significant increase in mass.

The p-value of the comparison between the control group and the group with applied drug is 0.0408793, indicating moderate significance.

When comparing the values from the Tab.3.5, we can see that the mean and median values of the cnt group are twice as large as the gdc group. Even tho the differentiations are small, the results still suggest a potential use of this biomarker for treatment. Comparing the values in the Fig.3.3 does not show great differences, but it is clear that the gdc group values are on average smaller than the cnt groups.

Tab. 3.5: Table with the results of comparing the protrusion max mass of cell groups with the drug (gdc) and without the drug (cnt).

Group	Movies	Cells	Mean	SEM	Min	Max	Median	ANOVA
cnt	4	69	0.124	0.006	0.012	0.233	0.121	
gdc	6	64	0.072	0.005	0.017	0.175	0.063	*

Despite low statistical values, this biomarker classification was nearly 70% accurate, with the model correct allocation of 84 cell of 106 into their groups, as seen on

Tab.3.6

Tab. 3.6: Classification table of protrusion max mass.

Number of cells	TP	TN	FP	FN	Accuracy (%)
68	14	32	1	21	67.6

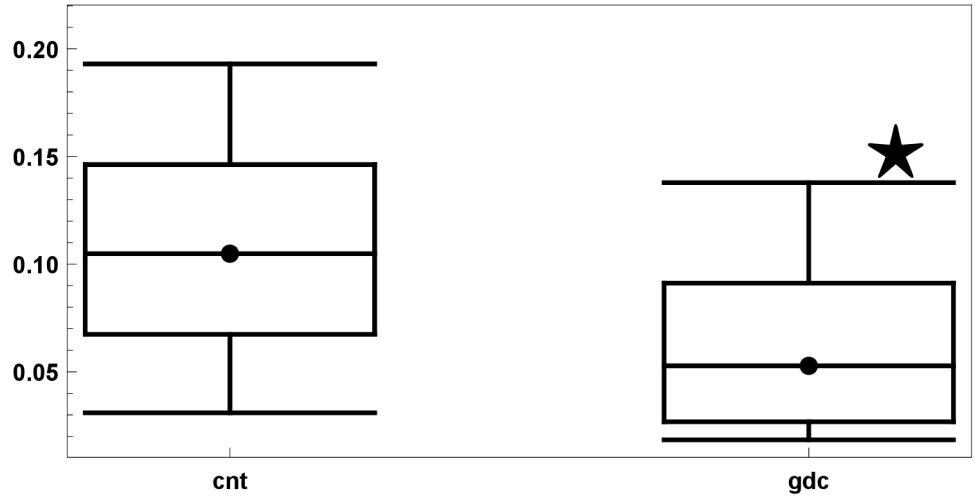


Fig. 3.3: Box and whisker plot of protrusion max mass displaying significance between control group and group with gdc, where dot represents median, box represents 50% of data and whiskers 80%.

3.4 Analysis of skew

The term skew describes the degree of distortion of values from normal distribution. In other words, it describes the degree to which the value differs from the normal distribution. In our case, we use skew as a distribution of mass intensities within a cell.

After applying the algorithm, we get the p-value 0.00801939, which indicates significance between the two groups. The skew value in our case is positive for both groups, which means that most values are less than the mean value. For the control group it is on average 0.56, which means that the cell value distribution is slightly symmetrical. Interesting is the value for the group with GDC, which is 1.05. This value describes that the distribution of cell values is very asymmetrical. Since the skew value indicates the intensity of mass within a cell, it is clear that the largest intensities will be found in the cell nucleus. Although the cancer cell divides more

rapidly, its membrane is constantly changing and growing, but it does not contain as much mass as in the nucleus. This may cause a fairly symmetric mass distribution in the cells of control group. On the other hand, in the group with GDC drug, cells have reduced motility and growth, causing a slow increase and variability of the membrane. This group therefore has an asymmetric mass distribution, since the larger portion of cell content is the nucleus itself.

As we can see on Tab.3.7 and Fig.3.4, skew value of gdc group is higher, meaning its distribution is more asymmetrical, than the cnt group.

Tab. 3.7: Table with the results of comparing the skew of cell groups with the drug (gdc) and without the drug (cnt).

Group	Movies	Cells	Mean	SEM	Min	Max	Median	ANOVA
cnt	4	69	0.56	0.039	0.069	1.272	0.504	
gdc	6	64	1.05	0.063	0.218	2.27	1.079	**

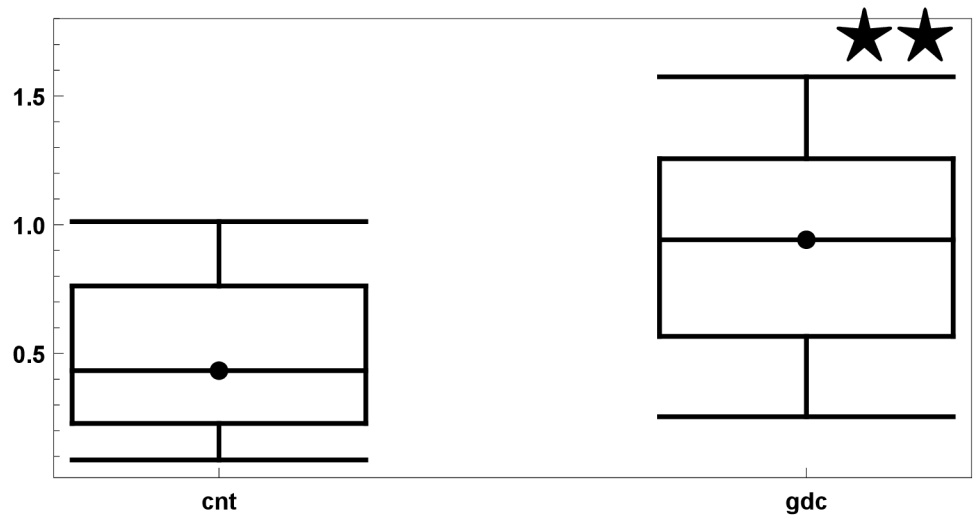


Fig. 3.4: Box and whisker plot of skew displaying significance between control group and group with gdc, where dot represents median, box represents 50% of data and whiskers 80%.

The SVM algorithm has also been used for this biomarker. As a result, the test was 72% accurate, as seen on Tab.3.8

Tab. 3.8: Classification table of skew.

Number of cells	TP	TN	FP	FN	Accuracy (%)
68	19	30	3	16	72

3.5 Analysis of shape factor - circularity

Shape factor describes certain shape properties of cells. In this case, the circularity of cells was calculated. From a physiological point of view, it is clear that cell structures will not have the shape of a precise circle, the value of 1. However, this biomarker might be of interest to look at because of the difference in shape between cancer cells and the cells whose growth and motility factors were suppressed.

Although the significance was not found by applying the algorithm, the p-value is 0.19354. Despite this statistical result, it can be seen in both the Tab.3.9 and the Fig.3.5 that the mean value of circularity is greater for cells of the GDC group than for cells in the control group. Because the value of the precise circle is 1, that is, the greater the value of the circularity, the more the object resembles the exact circle. In our case, neither group has approached this number, which is a physiologically normal phenomenon, since the cell does not have a precise circle shape. The reason why gdc group cells are "rounder" is the variability and irregular expansion of the cell membrane of cancer cells.

Tab. 3.9: Table with the results of comparing circularity of cell groups with the drug (gdc) and without the drug (cnt).

Group	Movies	Cells	Mean	SEM	Min	Max	Median	ANOVA
cnt	4	69	0.361	0.015	0.116	0.546	0.382	
gdc	6	64	0.407	0.017	0.086	0.567	0.437	

After application of the SVM algorithm, it was found that this biomarker is not suitable for machine learning recognition.

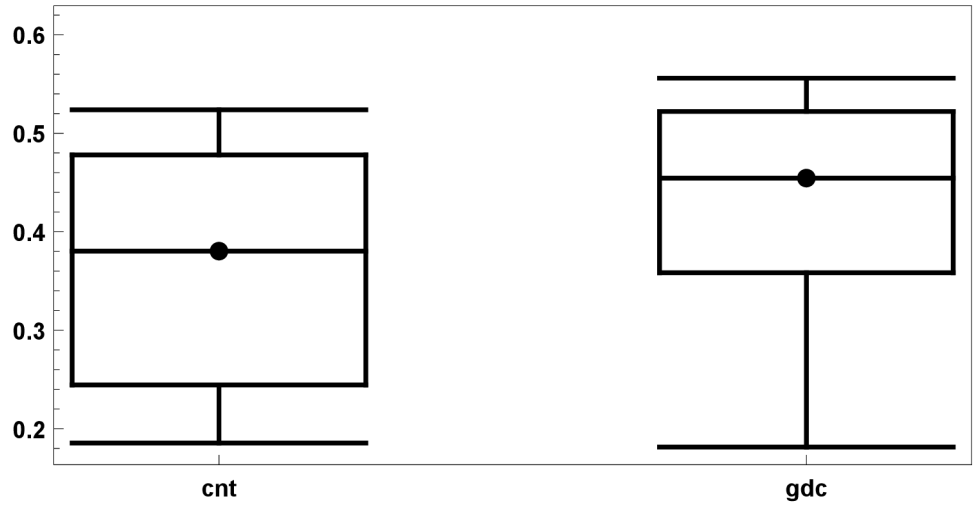


Fig. 3.5: Box and whisker plot of circularity displaying significance between control group and group with gdc, where dot represents median, box represents 50% of data and whiskers 80%.

3.6 Analysis of elongation

The last biomarker described in this work is also a shape feature of cell. It is elongation whose predicted value should be greater for cancer cells than for a group of cells with applied drug. This biomarker may be potential because of cell growth and movement. The only drawback is that elongation counts in a uniform direction, with cells, especially cancer cells, the movement is random.

Again, as in the previous biomarker, this one is not significant either. Despite these results, it can be seen again in the Tab.3.10 and in the Fig.3.6 that the elongation values are higher for the gdc group, confirming that elongation increases due to movement and growth.

Tab. 3.10: Table with the results of comparing elongation of cell groups with the drug (gdc) and without the drug (cnt).

Group	Movies	Cells	Mean	SEM	Min	Max	Median	ANOVA
cnt	4	69	1.813	0.077	1.078	3.114	1.650	
gdc	6	64	1.734	0.165	1.024	8.24	1.347	

Machine learning classification is not recommended for this biomarker, due to insufficient variability between groups, ie. values overlap, preventing proper machine learning. The cccuracy of this model is only 55%.

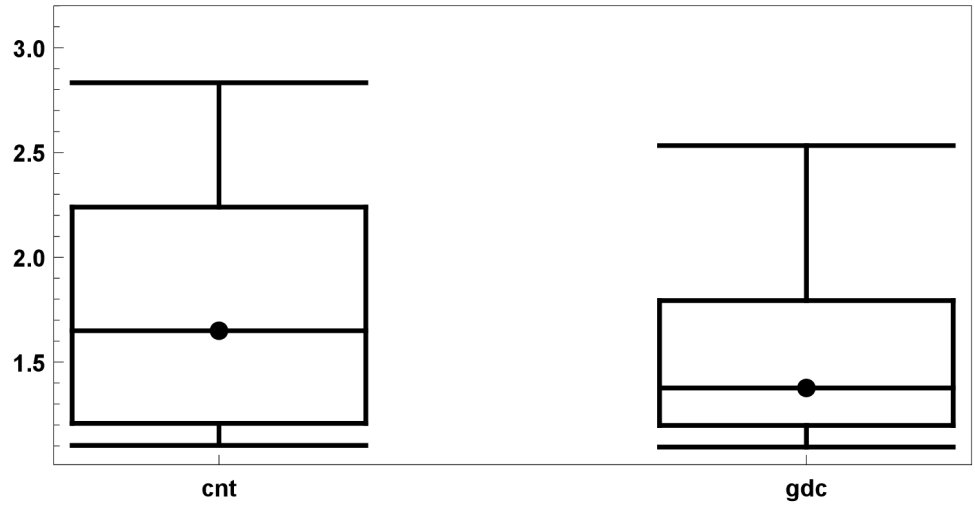


Fig. 3.6: Box and whisker plot of elongation displaying significance between control group and group with gdc, where dot represents median, box represents 50% of data and whiskers 80%.

Tab. 3.11: Classification table of elongation.

Number of cells	TP	TN	FP	FN	Accuracy (%)
68	14	24	9	21	55.88

3.7 Application of SVM to groups of biomarkers

In previous results in individual biomarkers, the accuracy of the classification was calculated for that group only. In this section, the SVM classifier will be trained by a combination of parameters and also tested for these combinations. The training data used was the same as in the individual biomarkers, 28 training cells for the GDC group and 33 training cells for the control group. All cells were used as test data, ie. 132 test cells.

3.7.1 Classification of two biomarkers

The first group used for classification consisted of protrusion area and protrusion max mass biomarkers. The reason for choosing these two was their significance in the ANOVA calculation. As we can see in the Tab.3.12, the accuracy of classification of 132 cell was 74%. If we compare this value with the individual biomarker values, we can see that the combination of the two biomarkers increased the accuracy of the model compared to the protrusion max mass alone, but decreased compared to the protrusion area.

Tab. 3.12: Classification table of protrusion area and protrusion max mass biomarkers.

Number of cells	TP	TN	FP	FN	Accuracy (%)
132	38	60	9	25	74

3.7.2 Classification of three biomarkers

Another group used for classification consisted of protrusion area, protrusion max mass and speed biomarkers. The speed has been added to compare how much this biomarker will affect training and classification of the model itself. As can be seen in Tab.3.13, the classification accuracy increased up to 78.79% after the addition of this biomarker. The reason for the improvement is that the speed itself has already a high classification accuracy.

Tab. 3.13: Classification table of protrusion area, protrusion max mass and speed biomarkers.

Number of cells	TP	TN	FP	FN	Accuracy (%)
132	56	48	21	7	78.79

3.7.3 Classification of four biomarkers

Subsequently, a group consisting of all significant biomarkers was classified, using protrusion area, protrusion max mass, speed and skew. The accuracy of classificationis 87.88%, as can be seen in the Tab.3.14. The reason for the increase in accuracy compared to the previous one is due to the skew biomarker.

Tab. 3.14: Classification table of protrusion area, protrusion max mass, speed and skew biomarkers.

Number of cells	TP	TN	FP	FN	Accuracy (%)
132	55	61	8	8	87.88

3.7.4 Classification of all biomarkers

The last group subjected to classifications consisted of all parameters. The classification accuracy was 84.85% as seen on Tab.3.15. If we compare this value with the

previous calculated, we find that it has decreased by nearly 3%. This decrease is due to the application of non-significant biomarkers such as circularity and elongation.

Tab. 3.15: Classification table of all parameters.

Number of cells	TP	TN	FP	FN	Accuracy (%)
132	55	57	12	8	84.85

3.8 Analysis using PCA

The basic usage of the PCA in this work was to reduce the data dimensions and to find out which parameter has the largest dispersion of values, in other words, which parameter among the others was best detected. Another reason for using it was to try to create new biomarkers by combining two parameters. The principal components of this method helped to achieve this. After data standardization, the covariance matrix was calculated for all parameters and thus the eigenvalues indicating the significance of the individual components. As mentioned, eigenvalues are sorted by size from largest to smallest, and the largest value is the first principal component. The result was then normalized to show the percentage of individual principal components. As can be seen in the Tab.3.16, the first principal component affects the result to 50%, according to the results in Tab.3.17 it can be stated that this component is formed by the variability of the protrusion area parameter. This is followed by a second principal component with 22.6%. The last principal component affects the result by only 2%.

Tab. 3.16: Percentage representation of individual components.

1st	2nd	3rd	4th	5th	6th
48.6	23.7	13.3	8.7	3.4	2.23

Tab. 3.17: First eigenvector of all parameters.

Speed	Protrusion Area	Protrusion Max Mass	Elongation	Circularity	Skew
0.46	0.52	0.45	0.17	-0.4	-0.35

Subsequently, graphs were created for better imagination of variability of individual parameters. Together we have 6 parameters that have been combined with

each other. As a result, 15 graphs of pairs of parameters were created to show how they are related.

We also calculated the percentage of individual principal components for the pairs of parameters. For example, if we take the protrusion area parameter and the circularity, both in the results and in Fig.3.9(c) it can be seen that the protrusion area has the most variability, which is the first principal component of the pair. The percentage of the first principal component is 80% and the remaining 20% is the second principal component. To create a new biomarker these principal components can be used in the equation

$$\text{new biomarker} = \mathbf{PC1} \times X + \mathbf{PC2} \times Y \tag{3.2}$$

where PC1 means first principal component, PC2 means second principal component and X,Y means individual parameters.

PCA was also applied to all parameters. This method has reduced the number of data and rotated the data distribution on the new x and y axes, which represent the first and second principal components. In Fig.3.7 we can see a new layout where two groups of data are plotted. The all figures, the orange color represents the GDC group cells and the blue color indicates the control group cells.

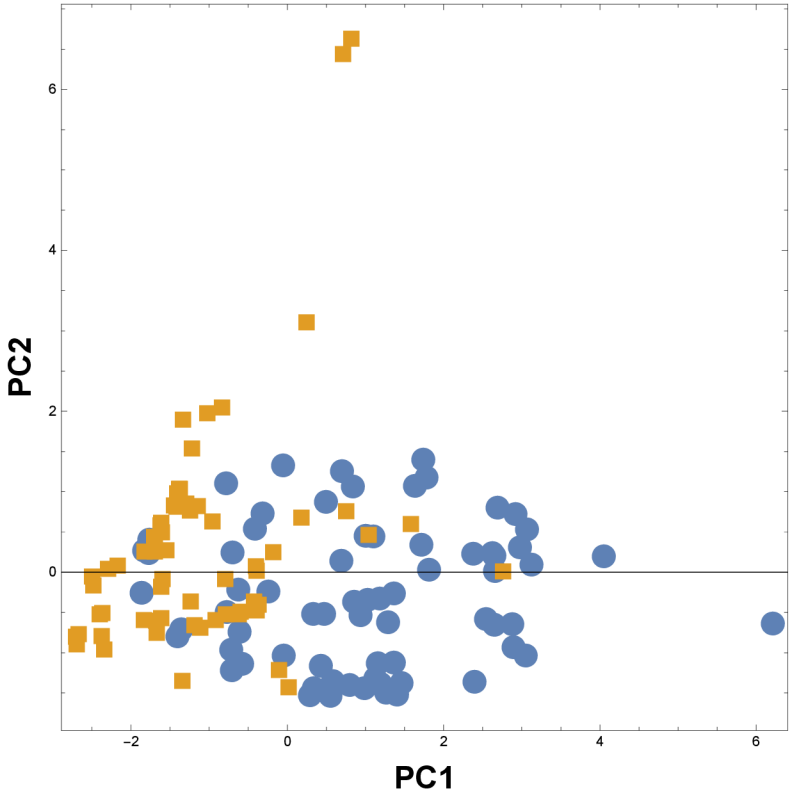
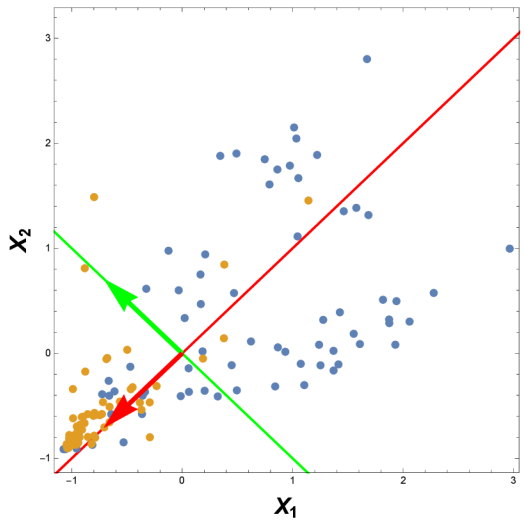
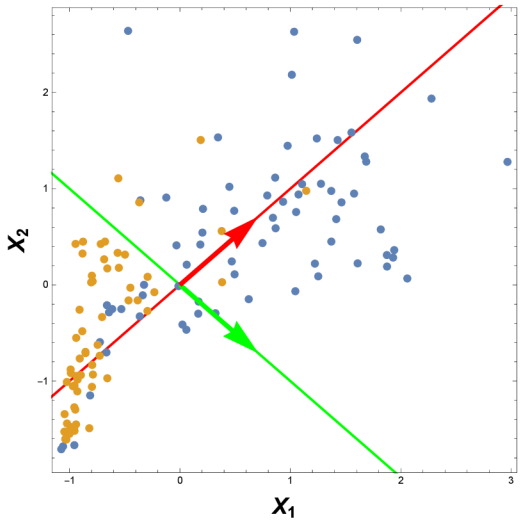


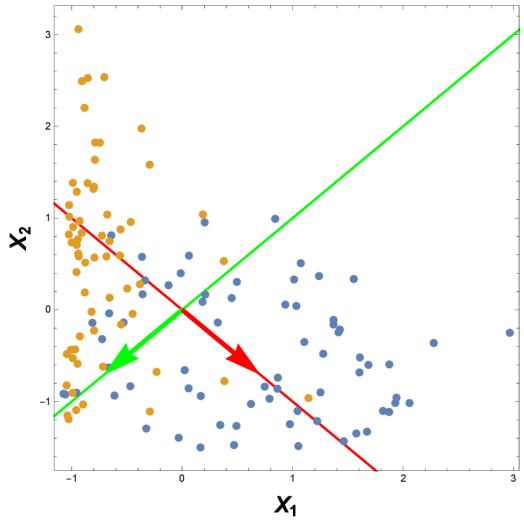
Fig. 3.7: Biplot of PCA applied to all parameters.



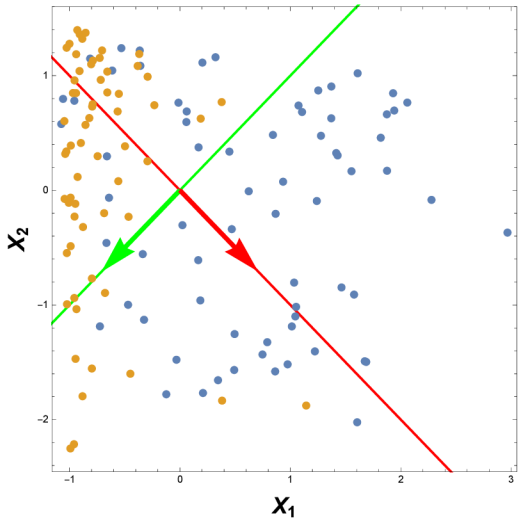
(a) Speed and protrusion area.



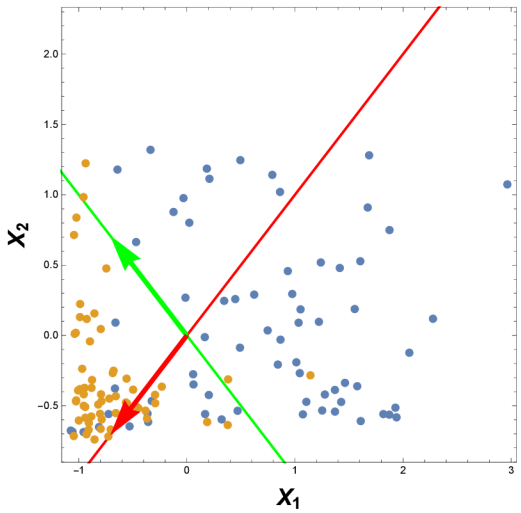
(b) Speed and protrusion max mass.



(c) Speed and skew.

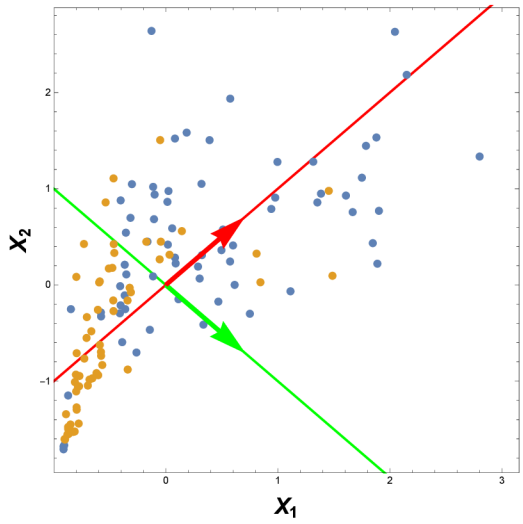


(d) Speed and circularity.

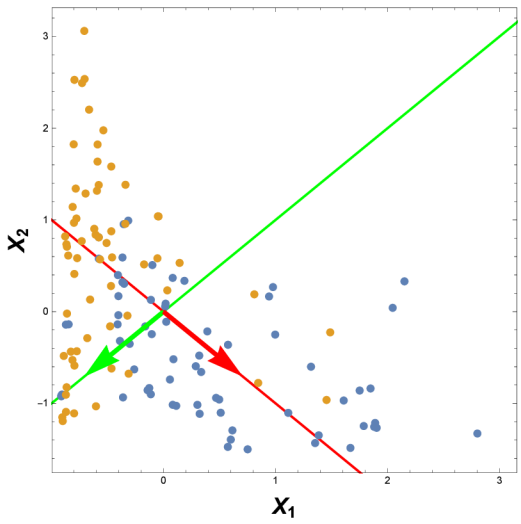


(e) Speed and elongation.

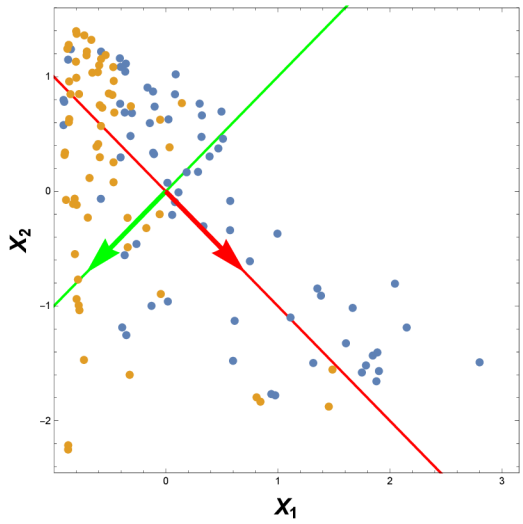
Fig. 3.8: Application of PCA on speed parameter(X1) with others(X2), where speed has green arrow.



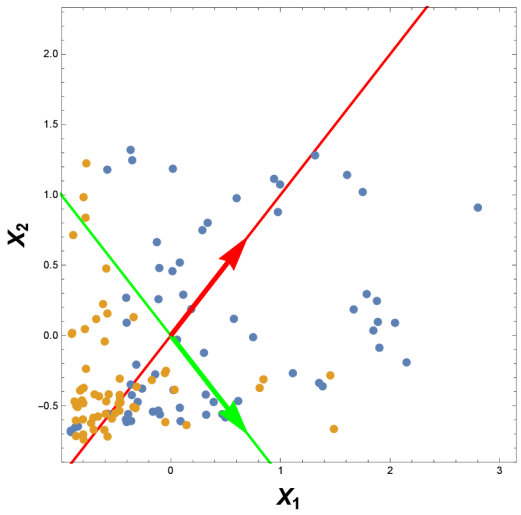
(a) Protrusion area and protrusion max mass.



(b) Protrusion area and skew.

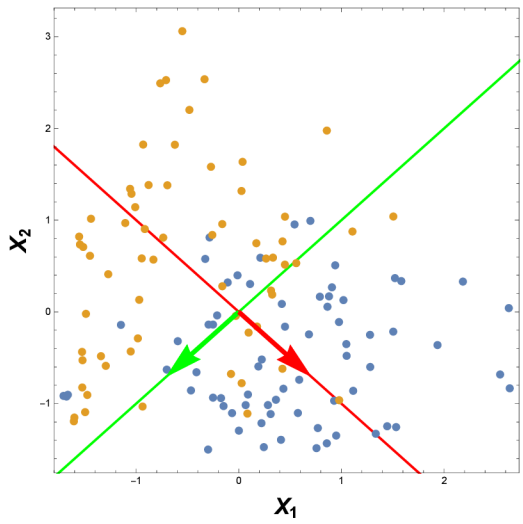


(c) Protrusion area and circularity.

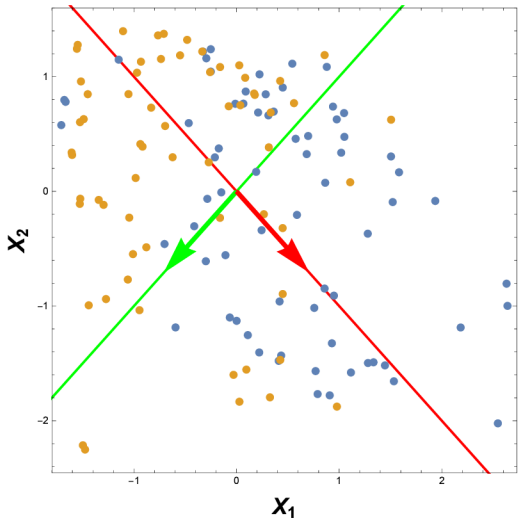


(d) Protrusion area and elongation.

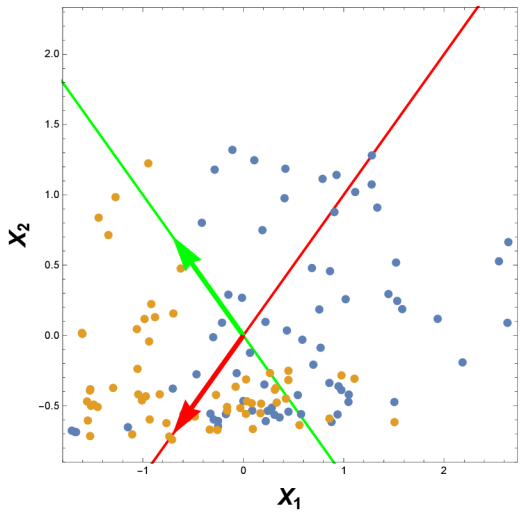
Fig. 3.9: Application of PCA on protrusion area parameter(X1) with others(X2), where protrusion area has green arrow.



(a) Protrusion max mass and skew.

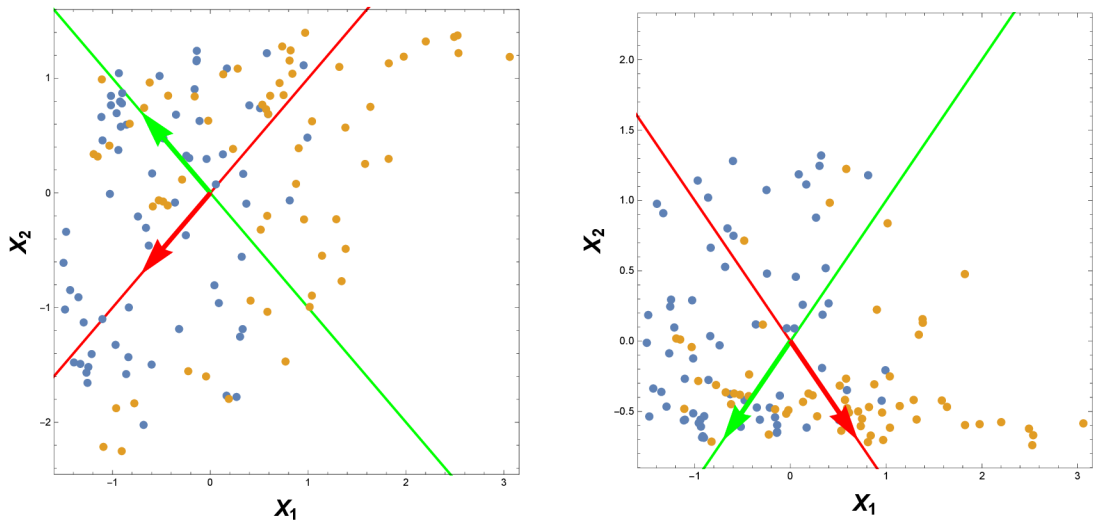


(b) Protrusion max mass and circularity.



(c) Protrusion max mass and elongation.

Fig. 3.10: Application of PCA on protrusion max mass parameter(X1) with others(X2), where protrusion max mass has green arrow.



(a) Skew and circularity.

(b) Skew and elongation.

Fig. 3.11: Application of PCA on skew parameter(X1) with others(X2), where skew has green arrow.

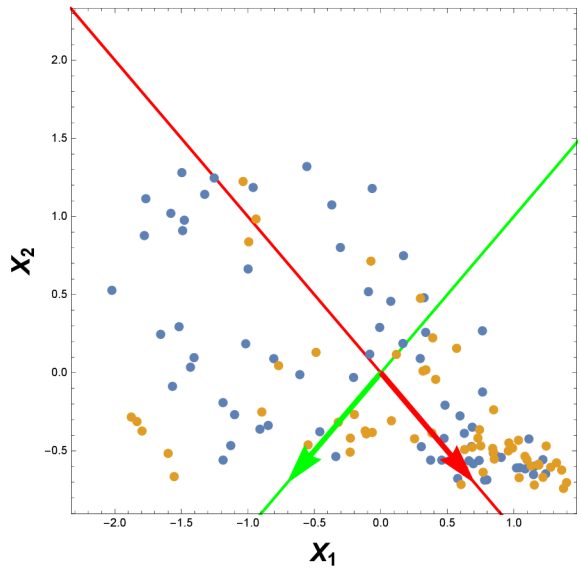


Fig. 3.12: Application of PCA on circularity parameter(X1) with elongation(X2), where circularity has green arrow.

4 Discussion

This work was focused on novel cancer biomarkers that could be used in the future for personalized cancer treatment. The issue of cancer is a widespread topic in the world, and in many cases the disease ends with the death of the patient. Therefore, it is important to focus on the core of the cancer, which are the cells. The aim of the work is to observe certain cell properties through quantitative imaging, making it easier in the future to identify drugs that respond well to a given type of cancer. This step may help with individual patient treatment. These biomarkers were observed between two groups of cells, one of which was control, where were the cancer cells, and the other consisted of the same group of cells, but with applied drug, an enzyme suppressor, which is in charge of many cell properties.

The first observed feature was speed. The reason for this choice lies in the physiological properties of the cells. It is generally known that cancer cells do not respond to the cellular factors responsible for the proper control of the cell and its growth. As a result, the cancer cell has uncontrolled growth, which in terms of speed means that it is increased over healthy cells. After comparing this feature between cell groups, results were obtained that described a high significance between the speed of both groups. These results have confirmed the theoretical assumptions as to the suitability of this biomarker.

Another characteristic feature of the cells is its protrusions. According to theoretical information, the variability of cell protrusions should be directly proportional to speed and growth. In this case, this again means that the values of the protrusions should be greater in cancer cells than in healthy ones. These protrusions themselves were the subject of two of the potential biomarkers of interest. The first was the protrusion area. The theoretical assumptions as to the suitability of this biomarker are that new protrusions arise and grow due to increased cell growth and motility. As a result of their growth, their area is also increasing in proportion. After applying an algorithm for detecting the area of protrusions on both groups of cells, it was found that the area of the protrusions is indeed much smaller in the group with applied drug.

Another observed parameter of the protrusions was their maximum mass. The assumption is the same as for the area, so the more the protrusion grows, the more mass there is. As a result, it was found that the maximum mass is not as significant as the area, but it is still possible to recognize the cancer cell from others. Thus, both biomarkers are potentially useful in the future for observing changes within cell groups.

A parameter that is quite interesting to interpret is skew. Generally, in statistics, this term describes the degree of distortion of values from normal distribution.

According to this parameter, we can find out whether the values are symmetrical to one another or there is a deviation of values. If asymmetries occur, it is possible to determine in which direction. Skew can therefore be positive or negative. If positive, mean is greater than median, which means that the "tail" points to larger values, but most values are less than mean. If it is negative, mean is smaller than median, which means that the "tail" points to lower values and most values are greater then mean . Also, the numeric value of skew is important. If equal to 0, the values are symmetrically distributed. If the value is greater than 0 but less than 0.5, the values are almost symmetrical. Above 0.5, the values are asymmetrically distributed.

Observing the value distribution is possible for different parameters. In our case, we chose to observe the mass intensities within the cell. As assumption, we expected more asymmetric values for drug-injected cells. The reason is again the nature of cell growth. If we want to interpret the distributed values, we need to find out where the greatest mass intensity is. We assume it is in the cell nucleus. The lowest mass intensity values are assumed at the cell edges. We also expect that due to the increased rate and growth of the cell, the ratio of the size of the nucleus to the size of the whole cell changes. Particularly, the distribution of values in the cancer cell should vary according to the size of the cells. In other words, as the cell grows, a small number of large intensities in nucleus will be suppressed by a large number of small intensities of cell edge. Indeed, after the skew was computed in both groups of cells, it was indeed found that cancer cells have an asymmetric distribution of intensities, but nearly two times less than the distribution of the cell group with applied drug. This result confirmed the assumption of future use of biomarker.

Another interesting feature of the cells is their shape. It goes without saying that, due to the growth and movement of cells, its shape changes as well. The cell shape features could show how much the shape changes and whether it is an indicator of the physiological state of the cell. From the point of view of shape, we discussed the circularity and elongation of the cell in this work.

It is generally known that the cell is not circular in shape. Despite this fact, we have observed how growth and motility affect the cell's circularity. Circularity of the exact circle is 1. We assume that due to excessive growth and random motility of cancer cells, the value of circularity will be very low. Although drug-injected cells have both slowed growth and motility, but their shape may be random, so it cannot be claimed that their shape will approach the exact circle. Nevertheless, we expect a circularity value for this group greater than for a group of cancer cells.

After finding the value of circularity of the two groups, it was confirmed that the drug group had greater value of circularity than the other. However, the values are not very different, so it is not possible to confirm the suitability of the biomarker.

As already mentioned, the other cell property observed was its elongation. A

prerequisite for suitability of this feature was again the factor of cells motility. It was assumed that the faster the cell moved, the greater the elongation would be. The problem with this parameter is that the direction in which the cell moves is purely random but the elongation is dependent on the direction of movement. Despite this, after calculating the elongation values for both groups, it was found that the elongation value is greater in the cancer cell group. The differences between the groups did not differ much again, so it is not recommended to mark this biomarker as suitable for individual treatment. By summarizing both cell shape properties, it can be stated that the shape of the cell is not a good indicator of the differences between a healthy and a cancer cell.

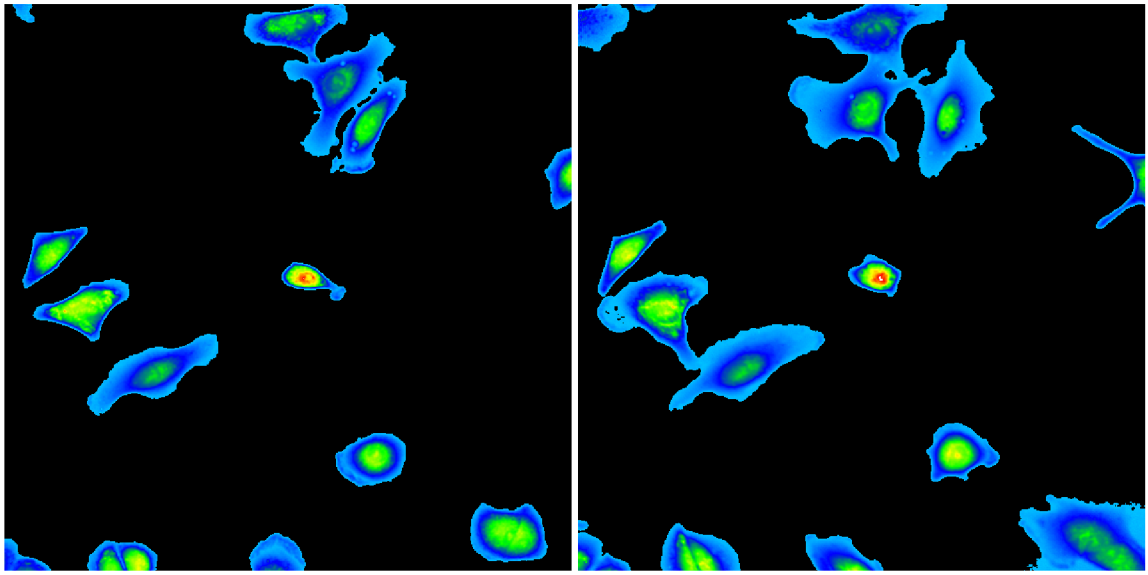
After assessing all these parameters, I would like to point to a Fig.4.1 , where the motility and growth of the control group of cells can be seen and the Fig.4.2 showing the motility and growth of the group with the applied drug. If we compare these two figures, it is indeed possible to see that cancer cells have much increased motility and growth as the second group. It is precisely those biomarkers that have been significant and potentially useful in the future for individual patient treatment are either dependent on growth or motility, or even on both cell parameters.

The next step was to apply the PCA method to determine which parameter is the most significant. According to the results, we assumed that the method will be most influenced by speed and least by circularity. This conjecture was supported by the fact that the variability of data, ie. the dispersion of data in individual parameters, is used in the PCA. Since the variability of speed was large and of the circularity was very small, the assumption was confirmed after PCA application on all parameters. Prospectively, it would be possible to create new biomarkers by the PCA method, a combination of two known biomarkers. It is possible that such biomarkers would find widespread use in individual treatments, but this is a future issue.

After evaluating all the results, it can be said that several factors can influence research accuracy. One of them is the application of all parameters to only two groups of cells. It would be interesting to apply all parameters to a group of healthy cells, where it would be possible to compare the results of cell groups with cancer, with applied drug and with healthy cells.

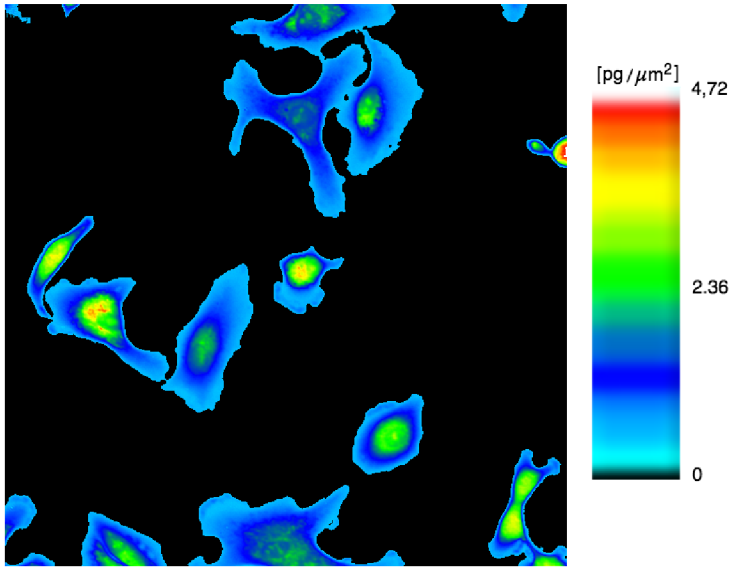
Another factor of inaccuracy may be the fact that it is not possible to determine exactly whether all cells of the cancer group really equally severe cancer transformation, which may create inaccuracies in the results of this group. Even though most cells are transformed and a small percentage may not be, this part of the group will still push average results.

The next factor can be the actual recording of the cells with a microscope, where each microscope has its own settings, which could in some way affect the results.



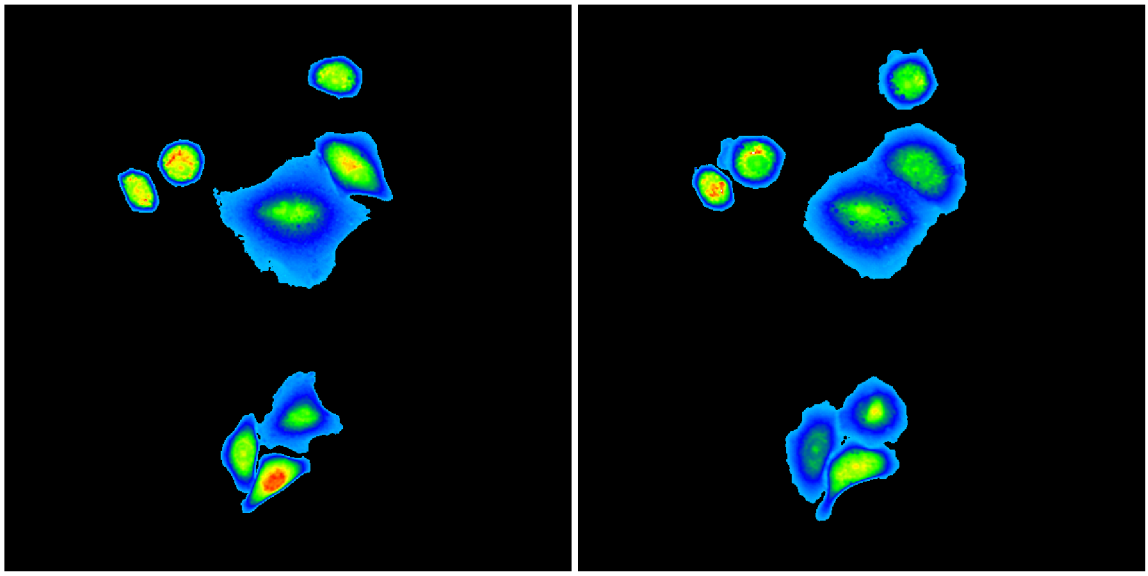
(a) First image of the cells from the movie.

(b) Cells after 10 hours.



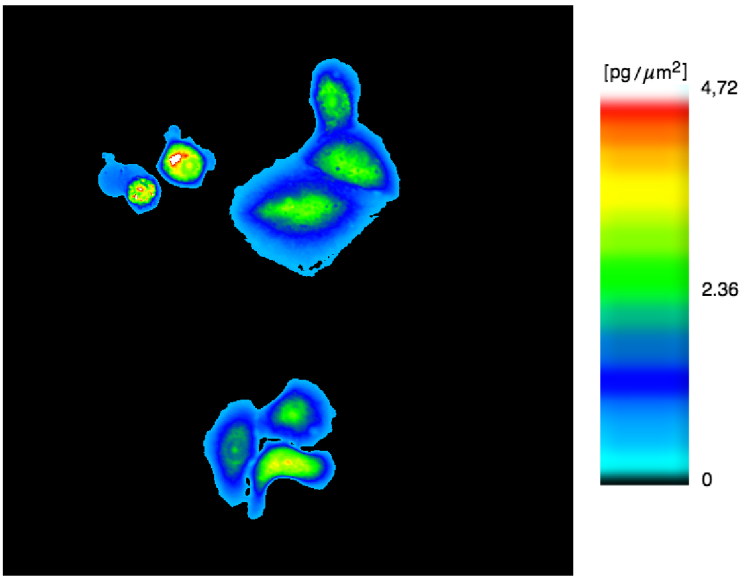
(c) Cells after 20 hours.

Fig. 4.1: Cellular motility of the control group of cells during the time-lapse recording.



(a) First image of the cells from the movie.

(b) Cells after 10 hours.



(c) Cells after 20 hours.

Fig. 4.2: Cellular motility of group of cells with drug during the time-lapse recording.

5 Conclusion

This work was focused on cancer biomarkers, which can be used in personalized treatment.

The first chapter briefly summarized the issue of cancer and the importance of further research to suppress its morbidity and mortality. An increase in incidence and mortality can be seen in the statistical results at the end of the chapter.

The second chapter is devoted to methods of analysis of cells and their structures. The principles of light microscopy are explained in more detail, where the QPI method and its advantages in cell analysis are also described. Image processing is briefly described including segmentation, tracking, feature extraction, and classification. The ANOVA statistical method, which is also used in the practical part, is described to improve the evaluation of the results. Another method described in this chapter was the principal component analysis, which was used in research to determine the significance of each parameter.

The third chapter describes the practical part in which the cell properties are analyzed as potential biomarkers. Altogether, 6 biomarkers were calculated, the most significant results being shown by biomarkers that are dependent on cell motility. The most significant and most accurate biomarker in the classification was speed. However, when monitoring cells from a biopsy, it is not advisable to use the speed due to cell aggregation. This problem is well handled by protrusion area and protrusion max mass biomarkers, where the newly allocated cell space is computed. By combining these two biomarkers, a 74% success rate of classification was found, indicating a suitable potential for biopsy cell observation. At the end of this chapter, the individual biomarkers are compared with each other using PCA method, which could create new biomarkers in the future. This method was also applied to all parameters, causing data reduction and new data distribution due to the first and second principal component.

References

[1] Love Richard R., ed. *Manual of Clinical Oncology.* Sixth Edition. Germany: Springer-Verlag, 1994. ISBN 3-540-58193-6. 0-387-58193-6.

[2] WEINBERG, Robert A. *The biology of cancer.* Second edition. New York, 2014. ISBN 978-0-8153-4220-5.

[3] MURPHY, Gerald P., Walter LAWRENCE and Raymond E. LENHARD. *American Cancer Society Textbook of Clinical Oncology.* 2nd ed. United States of America: American Cancer Society, 1995. ISBN 0-944235-10-7.

[4] ULDRIJAN, Stjepan. *Úvod do nádorové biologie* Lecture. Brno: Masaryk University, Faculty of Medicine

[5] Devi, P. Uma. *Basics of carcinogenesis.* Health Adm., 2004, 17(1), 16-24.

[6] BOFFETTA, Paolo, Stefania BOCCIA and Carlo LA VECCHIA. *A quick guide to cancer epidemiology.* New York: Springer, 2014. SpringerBriefs in cancer research. ISBN 3319050672.

[7] PARKIN, Max, Freddie BRAY, J. FERLAY, Paola PISANI. *Global Cancer Statistics, 2002.* A Cancer Journals for Clinicians, 2005, 55(2), 74-108.

[8] TORRE, Lindsey A. *Global cancer statistics, 2012.* CA: A Cancer Journal for Clinicians, 2015, 65(2), 87-108.

[9] Global Cancer Observatory. *Global Cancer Observatory [online].* France: IARC, 2018 [cit. 2018-11-25]. Available at: http://gco.iarc.fr/

[10] NAVRÁTIL, Leoš. *Medicínská biofyzika.* Praha: Grada, 2005. ISBN 978-80-247-1152-2.

[11] HEJTMÁNEK, Milan. *Úvod do světelné mikroskopie: Určeno pro posl. lék. fak.* 3rd. Olomouc: Vydavatelství Univerzity Palackého, 1993. ISBN 80-706-7308-7.

[12] HRAZDIRA, Ivo a Vojtěch MORNSTEIN. *Lékařská biofyzika a přístrojová technika.* Brno: Neptun, 2001. ISBN 80-902-8961-4.

[13] MADIGAN, Michael T. *Brock biology of microorganisms.* Fourteenth edition. Boston: Pearson, 2015. ISBN 978-0-321-89739-8.

[14] STEPHENS, D. J. *Light Microscopy Techniques for Live Cell Imaging.* Science. 300(5616), 82-86. DOI: 10.1126/science.1082160. ISSN 00368075. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1082160

[15] PLÁŠEK, Jaromír a Josef REISCHIG. *Kontrast v optické mikroskopii.* Vesmír. 1995, 74(638). Available at: https://vesmir.cz/cz/casopis/archiv-casopisu/1995/cislo-11/kontrast-opticke-mikroskopii.html

[16] PLÁŠEK, Jaromír. *Nové metody optické mikroskopie.* Pokroky matematiky, fyziky a astronomie. 1996, 41(1), p 1-24. Available at: https://dml.cz/handle/10338.dmlcz/139719

[17] SHRIBAK, Michael. *Differential Interference Contrast Microscopy (DIC).* Biomedical Optical Phase Microscopy and Nanoscopy [online]. Elsevier, 2013, 2013, p. 19-42 [cit. 2018-12-11]. DOI: 10.1016/B978-0-12-415871-9.00002-8. ISBN 9780124158719. Available at: https://linkinghub.elsevier.com/retrieve/pii/B9780124158719000028

[18] VYMĚTALOVÁ, Veronika. *Biologie pro biomedicínské inženýrství.* Praha: Česká technika - nakladatelství ČVUT, 2008-. ISBN 978-80-01-04013-3.

[19] WOLF, Emil, ed. *Progress in optics.* Amsterdam: Elsevier, 2012. ISBN 978-0-44-459422-8.

[20] LEE, KyeoReh, Kyoohyun KIM, Jaehwang JUNG, et al. *Quantitative Phase Imaging Techniques for the Study of Cell Pathophysiology: From Principles to Applications.* Sensors. 2013, 13(4), 4170-4191. DOI: 10.3390/s130404170. ISSN 1424-8220. Available at: http://www.mdpi.com/1424-8220/13/4/4170

[21] PARK, YongKeun, Christian DEPEURSINGE a Gabriel POPESCU. *Quantitative phase imaging in biomedicine.* Nature Photonics [online]. 2018, 12(10), 578-589 [cit. 2018-12-12]. DOI: 10.1038/s41566-018-0253-x. ISSN 1749-4885. Available at: http://www.nature.com/articles/s41566-018-0253-x

[22] ACHARYA, Tinku a Ajoy K. RAY. *Image processing: principles and applications.* Hoboken, N.J.: John Wiley, 2005. ISBN 0-471-71998-6.

[23] CHEN, Shengyong, Mingzhu ZHAO, Guang WU, Chunyan YAO a Jianwei ZHANG. *Recent Advances in Morphological Cell Image Analysis.* Computational and Mathematical Methods in Medicine [online]. 2012, 2012, 1-10 [cit. 2018-12-13]. DOI: 10.1155/2012/101536. ISSN 1748-670X. Available at: http://www.hindawi.com/journals/cmmm/2012/101536/

[24] PINIDIYAARACHCHI, Amalka. *Digital Image Analysis of Cells: Applications in 2D, 3D and Time.* Uppsala, 2009. ISBN 978-91-554-7398-3. Dissertation. Uppsala University.

[25] JAN, Jiri. *Medical image processing, reconstruction, and restoration: concepts and methods.* Boca Raton, FL: Taylor & Francis, 2006. ISBN 978-0-8247-5849-3.

[26] SONKA, Milan, Vaclav Hlavac a Roger Boyle. *Image processing, analysis, and machine vision.* 3rd ed. Toronto: Thompson Learning, c2008. ISBN 9780495082521.

[27] JAHNE, Bernd. *Practical handbook on image processing for scientific applications.*Boca Raton, Fla.: CRC, c1997. ISBN 0-8493-8906-2.

[28] MEIJERING, Erik, Oleh Dzyubachyk a Ihor Smal. *Methods for Cell and Particle Tracking.* Imaging and Spectroscopic Analysis of Living Cells - Optical and Spectroscopic Techniques [online]. Elsevier, 2012, 2012, p. 183-200 [cit. 2018-12-15]. Methods in Enzymology. DOI: 10.1016/B978-0-12-391857-4.00009-4. ISBN 9780123918574. Available at: https://linkinghub.elsevier.com/retrieve/pii/B9780123918574000094

[29] MEIJERING, Erik, Oleh DZYUBACHYK, Ihor SMAL a Wiggert A. VAN CAPPELLEN. *Tracking in cell and developmental biology.* Seminars in Cell & Developmental Biology [online]. 2009, 20(8), 894-902 [cit. 2018-12-15]. DOI: 10.1016/j.semcdb.2009.07.004. ISSN 10849521. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1084952109001517

[30] ULMAN, Vladimir, Martin Maska, Klas E G Magnusson, et al. An objective comparison of cell-tracking algorithms. Nature Methods [online]. 2017, 14(12), 1141-1152 [cit. 2018-12-15]. DOI: 10.1038/nmeth.4473. ISSN 1548-7091. Available at: http://www.nature.com/doifinder/10.1038/nmeth.4473

[31] BHATIA, Nitin. *Survey of Nearest Neighbor Techniques.* International Journal of Computer Science and Information Security. 2010, 8(2), 302-305. ISSN 1947-5500.

[32] COUPRIE, C, L Grady, L Najman a H Talbor. *Power Watershed: A Unifying Graph-Based Optimization Framework.* IEEE Transactions on Pattern Analysis and Machine Intelligence [online]. 2011, 33(7), 1384-1399 [cit. 2018-12-15]. DOI: 10.1109/TPAMI.2010.200. ISSN 0162-8828. Available at: http://ieeexplore.ieee.org/document/5639015/

[33] KANADE, Takeo, Zhaozheng YIN, Ryoma BISE, Seungil HUH, Sungeun EOM, Michael F. SANDBOTHE a Mei CHEN. *Cell image analysis: Algorithms,*

system and applications. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV) [online]. IEEE, 2011, 2011, s. 374-381 [cit. 2018-12-15]. DOI: 10.1109/WACV.2011.5711528. ISBN 978-1-4244-9496-5. Available at: http://ieeexplore.ieee.org/document/5711528/

[34] LINLIN, Zhu, Fan BAOJIE, Li BENJIN a Tang YANDONG. *A Hybrid Tracking Method Based on Active Contour and Mean Shift Algorithm.* In: 2009 Second International Conference on Intelligent Networks and Intelligent Systems [online]. IEEE, 2009, 2009, s. 70-73 [cit. 2018-12-15]. DOI: 10.1109/ICINIS.2009.27. ISBN 978-1-4244-5557-7. Available at: http://ieeexplore.ieee.org/document/5365679/

[35] BANKMAN, I. N. *Handbook of medical imaging: processing and analysis.* San Diego, CA: Academic Press, c2000. ISBN 0-12-077790-8.

[36] FRIEL,J. J. *Practical guide to image analysis.* Materials Park, Ohio: ASM International, c2000. ISBN 0871706881.

[37] KUMAR, Vijay a Priyanka GUPTA. *Importance of Statistical Measures in Digital Image Processing.* International Journal of Emerging Technology and Advanced Engineering. 2012, 2(8), 56-62. ISSN 2250-2459.

[38] *Roughness (3D) parameter: Height Parameters (peaks and valleys)* [online]. [cit. 2019-05-22]. Available at: https://www.olympus-ims.com/en/knowledge/metrology/roughness/3d_parameter/

[39] LI, Heng, Zhiwen Liu, Fengqian Pang a Yonggang Shi. *Cell dynamic morphology analysis by deep convolutional features.* In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [online]. IEEE, 2017, 2017, s. 2700-2703 [cit. 2018-12-25]. DOI: 10.1109/EMBC.2017.8037414. ISBN 978-1-5090-2809-2. Available at: https://ieeexplore.ieee.org/document/8037414/

[40] MELOUN, M., Militký, J *Metoda hlavních komponent v laboratorní praxi.* [online] [cit. 2019-05-09], available at: http://meloun.upce.cz/docs/publication/127b.pdf

[41] RINGNÉR, Markus. *What is principal component analysis?.* Nature Biotechnology [online]. 2008, 26(3), 303-304 [cit. 2019-05-09]. DOI: 10.1038/nbt0308-303. ISSN 1087-0156. available at: http://www.nature.com/articles/nbt0308-303

[42] SMITH, L. I. *A tutorial on principal components analysis.* 2002. [cit. 2019-05-10], available at: https://ourarchive.otago.ac.nz/bitstream/handle/10523/7534/OUCS-2002-12.pdf

[43] GANESHKUMAR, P. *Statistical methods in medical research.* International Journal of Clinical Trials [online]. 2014, , 1-2 [cit. 2018-12-25]. DOI: 10.5455/2349-3259.ijct20140501. ISSN 2349-3240. Available at: http://www.ijclinicaltrials.com/index.php/ijct/article/view/31

[44] GELMAN, Andrew. *Analysis of variance?why it is more important than ever.* The Annals of Statistics [online]. 2005, 33(1), 1-53 [cit. 2018-12-26]. DOI: 10.1214/009053604000001048. ISSN 0090-5364. Available at: http://projecteuclid.org/euclid.aos/1112967698

[45] SAWYER, Steven F. *Analysis of Variance: The Fundamental Concepts.* Journal of Manual & Manipulative Therapy [online]. 2013, 17(2), 27E-38E [cit. 2018-12-26]. DOI: 10.1179/jmt.2009.17.2.27E. ISSN 1066-9817. Available at: http://www.tandfonline.com/doi/full/10.1179/jmt.2009.17.2.27E

[46] LAMORTE, Wayne W. *The ANOVA Procedure.* Hypothesis Testing - Analysis of Variance (ANOVA) [online]. Boston: Boston University School of Public Health, 2018 [cit. 2018-12-27]. Available at: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTesting-ANOVA/BS704_HypothesisTesting-Anova3.html

[47] ZICHA, D at al. 1999. *TGFbeta1 Induces a Cell-Cycle-Dependent Increase in Motility of Epithelial Cells.* J Cell Sci 112: 447–54. http://www.ncbi.nlm.nih.gov/pubmed/9914157.

List of symbols, physical constants and abbreviations

ACC	Accuracy
ANN	Artificial Neural Network
ANOVA	Analysis of variability
DAPI	4',6-diamidino-2-phenylindole
df	Degrees of Freedom
DFT	Discrete Fourier Transform
DIC	Differential Interference Contrast
EBV	Epstein-Barr virus
FN	False Negative
FP	False Positive
GLOBOCAN	Global Cancer Observatory
HBV	Hepatitis B virus
HCV	Hepatitis C virus
HPV	Human papillomavirus
HTLV	Human T-lymphotropic B virus
IARC	International Agency for Research on Cancer
IR	Infrared
MS	Mean Squares
NN	Nearest Neighbor
P13K	phosphatidylinositol-4,5-bisphosphate 3-kinase
PCA	Principal Component Analysis
QPI	Quantitative Phase Imaging
SS	Sum of Squares
SVM	Support Vector Machines
TN	True Negative
TP	True Positive
UV	Ultraviolet
WHO	World Health Organization

List of appendices

A The contents of the enclosed DVD

The content of the enclosed DVD include an electronic version of the bachelor thesis, source codes and a cell data.

- `/` ..... root directory of the attached DVD
  - `documentation`
    - `bp_xplisk04` ..... electronic version of bachelor thesis
  - `source codes`
    - `CompleteBiomarkers.nb` ..... the main script
    - `analyseTrackDataRoutines.m` the package needed for initialization ANOVA functions
  - `data`
    - `2018_06_29_CntGdc.zip` ..... cell data