Silesian University of Technology in Gliwice, Poland
Automatic Control, Electronics, and Computer Science Department

Classification of white blood cells based on single-cell sequencing data for biodosimetry purposes

Katarzyna Sieradzka

Doctoral thesis

Supervisor: prof. dr hab. inż. Joanna Polańska
Advisor: dr Christophe Badie

Gliwice, 2023

Agenda

1 Doctoral dissertation motivation

Ionizing radiation is a ubiquitous factor that has always accompanied human life. Radiation exposure may bring adverse health effects, considering the absorbed radiation dose. Still, this factor has gained a lot in value due to the development of branches of medicine that currently widely use techniques for treating and preventing diseases utilizing the radiation phenomenon affecting the human body. The analysis of this factor's influence on the processes inside living organisms, particularly in the functioning and changes in individual cells of complex organisms, could be the key to minimizing the adverse effects of ionizing radiation in everyday life. Moreover, the possibility and necessity of conducting such considerations is provided not only by the aspect mentioned above of the ubiquity of this factor but also by the development of technologies that allow insight into the changes taking place at the cellular level. The need to study these aspects is all the more critical as more high-dimensional data is currently being generated, offering many analytical possibilities, but not supported by appropriate tools, and above all, by sufficient caution and prudence when processing complex datasets.

1.1 Theses of the doctoral dissertation

The presented doctoral dissertation is based on two fundamental hypotheses:

1. **Combining feature engineering methods and advanced dimensionality reduction techniques with unsupervised clustering algorithms allows for the efficient identification of white blood cell subtypes in single-cell RNA sequencing data.**
2. **The proposed intelligent and stratified algorithm of the training set construction supports the classification system, especially in the case of heterogeneous datasets.**

1.2 Objectives of the doctoral dissertation

This doctoral dissertation assumes two main goals of the single-cell sequencing experiments data analysis: **to create the analysis workflow necessary for recognizing 1 Gy dose irradiated white blood cells gene signature** and **to detect and separate white blood cell subpopulations** in the single-cell sequencing data. Additionally, one side goal is assumed: **the comparison of two machine learning-based methods in terms of control and irradiated cells classification and resulting irradiated cells' genetic profiles**. All these goals are affiliated with implementing many bioinformatics methods, especially the analysis workflow that automates specific analysis steps. The most important in terms of content and giving the most significant opportunities for subsequent manipulation of the processes of the applied analysis steps is the construction of an appropriate workflow related to the crucial stage of work, feature selection.

Detection of irradiated cells gene signature

Detection of the genetic profile of irradiated cells for high-dimensional data from single-cell sequencing experiments not only creates a vast field for applying many methods and innovative solutions. The analysis of such complex data carries the burden of analyzing very large in terms of dimensionality as well as the degree of complexity and possible hidden interactions or other properties of such data, which must be very carefully recognized before starting the proper analysis, leading to the achievement of the assumed goal. Failure to pay special attention to this aspect may lead to distortions in the results obtained and the lack of full explanation, and, in extreme cases, contradictions and incorrect interpretations of the results. The gene signature recognition of 1 Gy dose irradiated cells forces the implementation of the machine learning-based method that will be well-thought-out and

adjusted to the complexity of the task and the analyzed data. After introducing scRNA-seq white blood cell data, the created algorithm should learn the specific gene structures that allow for the most effective separation of normal and irradiated cells. Due to the significant complexity of the research problem, it is necessary to use a combination and cooperation of many tools and methods that determine the impact of individual genes and gene structures, including gene cooperation and joint combined influence on cell differentiation. One of the essential parts of the proposed workflow is the features selection procedure on detailed and iteratively carried out modeling methods. This process is often underestimated but extremely important. It affects the duration of subsequent analyses and, above all, the interpretability of the results. As an effect of the feature selection procedure, it is assumed that the list of selected genes will be created corresponding to the given problem, so the ionizing radiation factor influences gene expression differentiation. Thus, after applying this factor, the detected gene signature can be used to make a biological inference about the structure and strength of changes in white blood cells. It will also be essential to utilize the detected irradiated cells' genetic model in a broader and more general way, i.e., to distinguish control and irradiated cells derived from single-cell sequencing experiments for cell classification purposes. Conclusions from this work stage may benefit medical and industrial approaches.

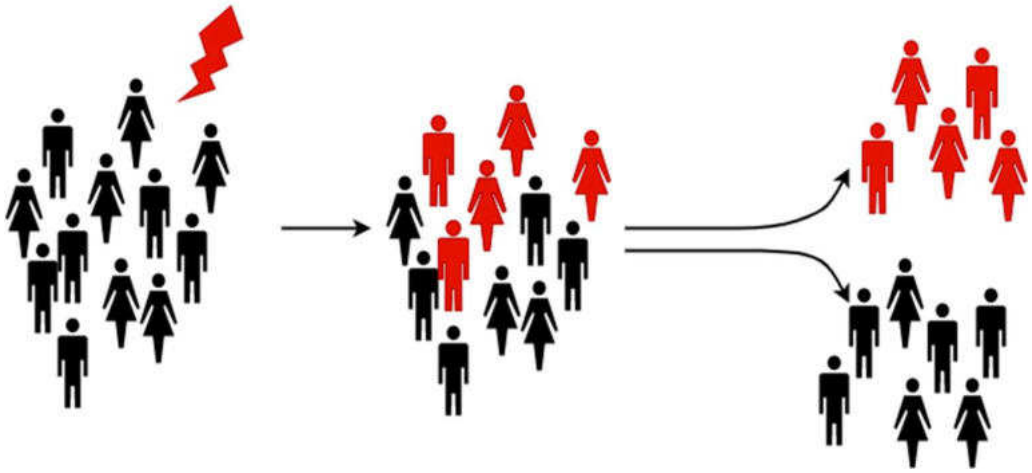


Figure 1. The aim of the doctoral dissertation is presented in a graphical form - the use of the genetic profile of irradiated cells for observation classification purposes.

Comparison of machine learning-based methods

The comparison of the application and the results obtained from two machine learning techniques, i.e., modeling based on logistic regression and neural networks, must address the two most important aspects of the work. The first is, of course, the stage of feature selection, as a result of which the gene signature of cells irradiated in an ex vivo environment will be determined. There is assumed that the signature for both approaches should be similar. However, the absolute number of features included in this signature is not considered, but their specific names and functions are. The expected similarities in gene composition are because only some of the genes present in the panel should show a specific response to irradiation. Moreover, the task of this work is to show these genes and exclude the others, which are not related to the radiation response but are responsible for several other functionalities, such as housekeeping genes, cell cycle genes, or genes responsible for the very differentiation of cells and their heterogeneity about the entire dataset. The diversity of the genetic composition of the signature may result primarily from the fact of prioritizing certain "behaviors" of genes (expression diversity or

cooperation of selected genes) due to internal algorithms and the diversity of the approaches used in this respect.

The second compared aspect, i.e., the generally described quality of the classification of control and irradiated cells in cooperation with the selected genetic profile, will make it possible to determine the usefulness of both solutions numerically and thus directly. It will also provide a broader view of the pros and cons of their use in the context of high-dimensional data. By design, the methodology woven into neural networks is much more complicated in terms of implementation than a simple learning algorithm based on logistic regression. The purpose of this stage of the work is to check whether any of these methods prevail in the case of the analyzed data, and if so, to what extent and in what context. The purpose of the comparison should be both the mentioned degree of complexity and the time necessary to carry out the calculations, but above all, what we can gain in the context of the clarity and explainability of the results, as well as the possibility of application and adaptation to a specific research problem and the specificity of the analyzed data.

Detection of white blood cell subpopulations

Recognition of cell subpopulations is another aspect directly related to different cell subtypes in the analyzed white blood cells. Such analysis of cell heterogeneity may open further comparative and analytical possibilities for the undertaken research. Differences in the responses of white blood cell subtypes may be significant enough to interfere with the self-learning classification algorithms. Due to this, existing subpopulations could significantly impact the classifiers' learning process results. Particular subtypes of white blood cells differ significantly not only in the case of their responses to stress factors such as ionizing radiation but also in terms of the sensitivity of such a response. Some subtypes of white blood cells are very sensitive to changes in the surrounding environment, which interacting even at low intensity may trigger apoptosis pathways. On the other hand, some of them, up to a certain threshold level, do not develop significant reactions to environmental changes. Due to such large fluctuations and subpopulation variability of individual white blood cell subtypes, it becomes an aspect that has a potential impact on the quality and accuracy of the results related to the genetic profile of irradiated cells. Classifier learning processes may focus on irrelevant aspects from the point of view of the problem under study and give the advantage on the described subpopulation relationships. Consequently, it may lead to the determination of a genetic profile containing interesting radiation response genes and some features responsible for the internal heterogeneity of the cell set. To avoid the distortion of the results, this aspect of differentiation in terms of subpopulations of white blood cells should be considered and treated as a priority. If such variability is detected, appropriate measures should be taken to prevent the influence of this factor on the obtained genetic profile of irradiated cells. In the detected genetic profile, there is no space for additional features that determine an aspect other than the response to ionizing radiation.

2 Introduction

Radiation is a naturally occurring factor on Earth that is an integral part of our lives. It means that our life has permanently moved, developed, and evolved in the presence of radiation. We come into contact with radiation not only consciously due to its omnipresence. We are exposed to even minimal doses of radiation every day. We have direct contact with radiation in our work, food, and water. Our houses are also a source of radiation, not only because of the more and more modern appliances for everyday use but also because of the material our houses are made of. Houses built of stone and brick have higher radiation levels than houses constructed of wood [1]. This shows that not only is what appears to be the apparent emission source but also several radiation sources that people need to be aware of. In the real world, many unconsciously selected environmental elements are potential emitters. The air we breathe to produce energy and live is also a significant radiation source. Of course, the level of radiation, unquestionably related to its effect on our bodies, is also essential. This level can vary significantly depending on where we work and live. Our organisms, tissues of skin, muscles, bones, and fluids inside our bodies also have naturally occurring radiating elements [1]. Analyzing the effects of ionizing radiation on the human body is a highly complicated process, often requiring many analytical tools and techniques.

A completely new single-cell sequencing method was beneficial for obtaining material for bioinformatics analysis. This method allows for simultaneously studying the genetic material of many cells, thus creating a pervasive and accurate set of information about the expression of individual genes [2] [3]. This method is much more accurate than its predecessor, the bulk analysis of genetic material, which averaged expression values over many cells [4] [5]. In the scRNA-seq research, it became possible to determine the expression (more precisely estimated count values) of each gene in the panel for individual cells, thus creating more extensive but highly informative sets of count matrices [6]. As a result of applying this technique, it has also become possible to automatically distinguish individual cells in the context of cell subpopulations or even to determine the difference in responses of individual cells to the applied external factor, such as ionizing radiation.

The challenge of the emerging scRNA-seq technique, more precisely of the multidimensional and complex count matrices generated, is effectively tackled in bioinformatics analysis. This kind of analysis is very progressive nowadays. After introducing the new method of sequencing, all the tools whose task was to meet the increasingly precise and more frequent challenges posed by biologists and doctors took place very quickly. To date, some tools and solutions have been developed for the quality control of data matrices, normalization of counts, or prediction based on a single-cell data set. Among the most popular tools are Partek Flow [7], Cellxgene [8], ROSALIND [9], and Cellenics [10]. These tools differ in the possibility of using data from different single-cell technologies and the range of possible steps in bioinformatics analysis. Apart from the technology in which the data was generated, their format is also essential. Not all tools work with raw data in fastq files, and not all support the already pre-processed count matrices. Several others support the transition processes between the stages of creating an input data structure suitable for analysis, such as Kallisto [11], STARsolo [12], or analysis protocols generated directly by sub-teams of companies dealing with single-cell data sequencing.

However, a large part of the tools do not answer the problem of many possible goals of the analysis; hence, there are still not many universal solutions that will fit each research problem perfectly [13]. Many existing tools often do not allow the user to select the particular steps of the study thoroughly or to omit the irrelevant steps in the context of the goal set. Hence, modifying existing or creating new tools is still necessary. For a scrupulous researcher, it is crucial to understand the individual steps of the tool's operation and to be able to interpret the results biologically. It is, therefore, essential to create a tool that will provide many possibilities regarding various research problems, as well as the fullest possible editability of individual parameters and steps of the analysis. This way, the tool will be able to create a uniform research structure, fully adapted to the data's needs and the research analysis's purpose.

2.1 Sources of radiation

Radiation sources can be divided into those that occur naturally and those that man has produced. The naturally occurring sources are cosmic radiation and radioactive minerals on the Earth (terrestrial) and the human body (internal).

Cosmic radiation is caused by highly energetic particles from the sun and stars entering the Earth's atmosphere. These particles are a direct source of radiation falling to the ground or indirect if it reacts with the atmosphere's components, influencing the type of radiation [14]. Suppose we consider the distribution of cosmic ray levels; of course, the closer to the source, the greater the dose. Therefore, the height above sea level of the place of work, residence, or stay significantly defines the received dose - the higher, the greater the received dose [15].

Internal and terrestrial radiation sources are related to the presence of radioactive minerals. This radiation is associated with the company and decays mainly of uranium, thorium, and radium [16]. The presence and decay of radioactive elements began before time began. Radioactive elements, their trace amounts, and decay products can be found everywhere in our environment [14]. These elements occur naturally in rocks, groundwater, and soil. Also, there is radon in the air, which directly affects living organisms, and contains radioactive carbon and potassium. However, the absorbed radiation dose differences may be significant for different regions. Not only is the terrain significant, but also the differences in uranium and thorium concentrations in soil in the various areas [17]. On the other hand, the general level of external radiation, considering terrestrial emission, is so low that its impact on human health is unlikely [17].

The last naturally occurring factor of radiation is a living organism, including the human body, which from birth to death, is characterized by the presence of both radioactive potassium-40 and carbon-14 [18]. Additionally, radioactive minerals (carbon, potassium, uranium, thorium, radium) can enter the body through food, drink, and air. The last mentioned case is crucial as most human exposure to natural radiation, especially radon, comes from breathing, the largest source of natural radiation [17]. The human body contains insignificant amounts of radioactive elements due to its continuous metabolism [14].

Radiation related to human presence and activities can be classified into two subcategories: radiation in public life and radiation linked to occupation [19]. Radiation in public life includes both sources of radiation coming from medical procedures and consumer products. Some medical procedures, one of the essential sources of radiation created by man, are related primarily to saving health and life and improving the quality of life. We have no direct influence on this type of emission, and the principle of improving patients' condition should undoubtedly be followed. However, it should be remembered that very often, undertaken treatment procedures are associated with exposure to significant radiation doses. For example, whole-body computed tomography (single design) exposes the patient to 10 mSv, while the annual dose of cosmic rays ingested at sea level is 0.3 mSv [14]. Radiation associated with consumer products comes from many areas of human life: building and road construction materials, combustible fuels (including gas and coal), X-ray security systems, the ceramics industry, tobacco, and smoke detection systems. People are also exposed to lower doses of radiation from the nuclear fuel industry, from extraction to disposal of spent fuel, and precipitation from atomic weapons testing and nuclear reactor accidents (Chernobyl) [19].

Today, many professions and jobs increase workers' exposure to radiation. The occupations requiring special attention are related to the fuel cycle, industrial radiography, radiological and nuclear medicine departments, atomic power plants, and research laboratories. Depending on the work performed and the radiation sources present, different exposures exist in different areas. On the other hand, due to the particular vulnerability to the negative impact of the radiation dose taken in the course of work, the workers of the exposed areas are under exceptional control of the organs whose task is to limit the exposure of adults working in the radioactive environment [19]. Instruments called dosimeters are often used to determine the degree of exposure. Table 1 is based on a report by the National Council on Radiation Protection & Measurements (NCRP) [20] and lists radiation doses from various sources.

Table 1. Radiation sources and their radiation dose.

Source of radiation	Radiation dose [mSv]
Whole body CT (single procedure)	10.00
Upper gastrointestinal X-ray (single procedure)	6.00
Head CT (single procedure)	2.00
Cosmic radiation – high elevation (annual)	0.80
Mammogram (single procedure)	0.42
Cosmic radiation – sea level (annual)	0.30
Radiation in the body (annual)	0.29
Terrestrial radioactivity (annual)	0.21
Chest X-ray (single procedure)	0.10
Leaving near a nuclear power station (annual)	<0.01

2.2 Types of radiation

As shown before, the presence of radiation in our lives is inevitable. Moreover, nowadays, radiation is utilized in many areas of life. Radiation surrounds us in everyday life, both in the environment of our life and at work, and when undertaking many medical procedures necessary to maintain health and life. It is essential to distinguish between two types of radiation: non-ionizing and ionizing. Additionally, ionizing radiation is divided into alpha particles, beta particles, gamma rays, and X-rays.

The difference between the two types of radiation is that non-ionizing radiation carries much smaller energy than ionizing radiation. This means that when non-ionizing radiation interacts with an atom, it does not cause ionization, i.e., it does not change the neutral charge of the atom to positive or negative. Examples of non-ionizing radiation are microwave waves, radio waves, visible light, laser light, and ultraviolet light. This radiation cannot ionize, but it can cause the atom in the molecule to move or even make it vibrate [1]. Due to this, some non-ionizing radiation can seriously damage the tissues of living organisms when the exposure is too high. A vivid example of the described case is sunburn or skin cancer, both arising from overexposure to UV radiation. On the other hand, ionizing radiation has enough energy to remove electrons from the atom, thus creating ions. This type of radiation, when used wisely, has excellent potential in both medicine and industry. Still, if used without prudence and the necessary protection against their operation, it can cause serious health problems, irreversible effects, and even in extreme cases, the death of a living organism.

Alpha particles are made of two protons and two neutrons from the nucleus. They are positively charged and heavy (the heaviest type of radiation particles); therefore, they do not travel far by air. Moreover, they cannot penetrate through clothing, paper, a thin layer of water, or even skin. Alpha particles are emitted by naturally occurring minerals such as uranium-238, thorium-232, rad-226, polonium-210, americium-241, and radon-222.

Beta particles are electrons formed from an atomic nucleus during radioactive decay and are not attached to the nucleus. They are negatively charged particles of low mass and can travel farther by air than alpha particles. Beta particles can penetrate the skin into the deepest layers of the epidermis, where new cells are produced. The clothing partially protects against this type of radiation. Instead, a sheet of plastic stops them. Examples of radioactive materials that emit beta particles are hydrogen-3 (tritium), carbon-14, phosphorus-32, sulfur-35, and strontium-90.

Gamma rays and X-rays are widely used in medicine nowadays. Contrary to the previously described radiation particles, they do not have any charge or mass and are therefore called photons. These types of radiation often accompany alpha and beta particles as well. They can carry an extensive range of energies. Gamma and X-rays pose a radiation hazard to the entire body because they penetrate most materials. They can travel a long way in the air and human tissues. To protect against gamma

rays, dense materials should be used. Lead or thick layers of concrete are used to stop this type of radiation. The difference between gamma rays and X-rays is that they come from different parts of the atom. X-rays are emitted from a process outside the nucleus on electron shells, while gamma rays come from inside the nucleus. Generally, X-rays contain less energy than gamma rays and have a lower penetrating force. The gamma-emitting radionuclides are technetium-99m (metastable), iodine-125, iodine-131, cobalt-57, cobalt-60, cesium-137, and radium-226.

2.3 Health effects of radiation

Even though judiciously and carefully used radiation brings undoubted benefits in many areas, such as medicine or industry. But if treated without due fear, it can cause irreversible damage to human organisms. All types of radiation carry energy that, when entering the tissues, can destroy cellular functions and change the genetic material. Of course, how the exposure will affect the body depends on three key factors the speed at which a specific dose of radiation was received, where the dose was absorbed, and the body's sensitivity to radiation. If the entire body is irradiated, a much more harmful effect can be expected than if only a specific part of the body is irradiated. The dose rate is a particular radiation dose delivered to the body within a specific time window. However, how a certain dose of radiation will affect the tissues depends mainly on the age and general condition of the body. The developing fetus is most exposed to the harmful effects of radiation. Young people, whose cell division and tissue development are rapid, are also at an increased risk. People with immune system defects and the elderly are also at increased risk [21].

Generally, the effect of radiation on a living organism and its tissues does not differ from that of other toxic substances. When a cell is destroyed, three main things can happen. The first is the complete repair of the changed cell and its return to normal functioning. Another is when the cell does not repair properly, which changes the cells' functions. This can lead to the development of neoplastic disease. The final aspect is cell death, which is not always the worst option. If a few misshapen cells die, the body will be fully functional again, and the changed cells will not pose a risk of cancer development. However, if too many cells within a given organ die, it may result in organ failure, which in critical situations will lead to the organism's death [21].

Irradiation exposure can be divided into three categories high, medium, and low. Exposure to radiation, such as nuclear weapons use, causes massive damage to cells, tissues, organs, and the entire organism. On the other hand, careful control of high doses of radiation can save lives. In cancer therapies, high doses of radiation are aimed directly at specific areas of cancer cells, which causes their destruction without the need to irradiate other parts of the body. High levels of radiation are doses above 1000 mSv [1]. The higher the radiation dose, the greater the chance of death. Due to the sudden death of many cells, high doses of radiation can trigger a violent response in the body, such as Acute Radiation Syndrome (ARS). It is a disease resulting from irradiating the whole body, or a large part of it, with high doses of radiation over a brief period. Sensitive to the harmful effects of radiation are primarily cells in the mitotic phase, where the genetic material is most exposed. Among the white blood cells, lymphocytes are the line most exposed to ionizing radiation, the first to be depleted in ARS [22].

Moderate exposure to radiation does not kill the exposed organism but can cause various cellular changes. The changed cells, dividing into new cells, can produce abnormal ones. The neoplastic process may begin with the appropriate accumulation of altered cells within a given tissue or organ. Exposure to moderate radiation levels can also cause changes in reproductive cells, which can be passed on and accumulated over the next generations [1].

Exposure to low doses of radiation comes primarily from the environment that affects the body, and these doses can also damage reproductive cells and lead to cancer development. They affect cells and tissues but do not immediately cause problems with the functioning of organs or the body. Moreover, low doses of ionizing radiation can even stimulate DNA repair processes [1].

The influence of particular levels of radiation on living organisms is intuitive. But it is also essential to know how different types of ionizing radiation can affect the body. Alpha particles are hazardous in

their effects because the ionizations they cause are very close to each other, so they can transfer all the stored energy to a limited number of cells. Beta particles have greater penetrating power than alpha particles, but the ionizations they cause are much more dispersed in space. Additionally, pollutants that emit beta particles that remain on the skin for long periods can cause severe skin damage. The last type of radiation, gamma radiation, thoroughly permeates the human body. In this way, it can cause ionizations that damage tissues and genetic material.

The complete list of recommendations and effects of ionizing radiation issued by the International Commission on Radiological Protection [23]. It describes in detail the biological effects of ionizing radiation depending on the dose taken, the issues of exposure to radiation in connection with medical procedures, and environmental facilities subject to special radiological protection.

2.4 Single-cell RNA-sequencing method

Single-cell RNA sequencing is a technique that extracts information about the expression of individual genes. A distinctive feature of this technique is that the expression information relates to a single, specific cell and not the averaged values for the set of analyzed cells, as in the previously utilized data acquisition methods. Such advanced and detailed technology allows for studying many aspects of biologists' dilemmas with unprecedented accuracy and many new possibilities that have become sincere thanks to scRNA-seq.

Currently used scRNA-seq protocols involve five steps: isolation of genetic material from single-cell RNA, reverse transcription, amplification, library building, and sequencing. The isolation of the research material is a crucial stage from the point of view of the subsequent sequencing steps, but it requires careful selection of viable cells. At this stage, several different possibilities appear, with the choice of techniques for isolating single cells or the appropriate indexing of single cells [24]. The separated cells must then undergo cell lysis so that the genetic material accumulated inside the cells becomes available and to capture as many RNA molecules as possible. In this step, polyadenylated mRNA determination using poly[T] primers is often used to avoid analyzing molecules other than those of interest. Then, as a result of reverse transcription, mRNA is transformed into complementary DNA molecules (cDNA). At this stage, various types of cDNA tags are often used. One is the unique molecular identifier (UMI), which uniquely identifies individual mRNA molecules recovered from one cell. Such molecule marking procedure is used before the amplification step as it allows the determination of read families of the same RNA fragment. After the PCR process, i.e., the process leading molecules marked with cDNA amplification, the collected material is combined and sequenced. Because, in theory, any eukaryotic cell can be analyzed using scRNA-seq [24], biomedical researchers have taken up the challenge of creating a transcriptomic atlas for each type of human cell called the Human Cell Atlas [25]. It was developed to create maps of different cell types through which we can describe and understand their complex functions and communication networks.

Due to the great interest of researchers in the scRNA-seq methods, explained by the great potential of this technique of biological data acquisition, many protocols have been developed, which include the subsequent and very detailed steps of the analysis. The creation of many instructions for scRNA-seq is related primarily to the type of analyzed cells and the study’s intended purpose. For example, poorly differentiated cell subtypes analysis may differ from analyzing the cellular response to a specific external factor. The protocols also vary in detail, for example, the minimum number of mRNA molecules necessary to determine the expression of a given gene. Therefore, some are more specific for poorly expressed genes, and the significance of this aspect entirely depends on the goal of biological data analysis.

Upon receipt of the raw sequencing data, the question arises, how to carry out further analysis steps to achieve the intended biologically interpretable goal? The answer to this question is not simple and unambiguous, mainly due to the sequencing data's rather complicated structure and the data analysis's specificity, where precision in the decisions made is essential. Platforms that perform sequencing come to the rescue, creating free applications for customers that enable fundamental data analysis. However,

in this case, and the possibility of generally available data analysis packages already available on many programming platforms, there is a problem with the 'black box' in which we do not know how the data is processed. Moreover, we do not have complete control over the subsequent stages of the analysis. Therefore we are wondering whether the data is analyzed in a way that will ensure the achievement of the set goal. There is a need for the work of specialized bioinformatics, who becomes a mediator between a biologist who expects biologically interpretable results and a programmer whose task is to implement appropriate software that gives a full range of analytical, control, and modification possibilities for each stage of the study. Until we can use tested algorithms and platforms that meet the requirements mentioned above, including complete adaptability to the data type and purpose, it is necessary to create analysis tools and protocols related to only a single study.

2.5 White Blood Cell subpopulations

The problem of cell subpopulation identification is not yet widely discussed in the literature, especially when considering the set of white blood cells. Looking from the perspective of the generality of the collection, in the context of the recognition of cell subpopulations in data from single-cell experiments, the most significant expenditure is currently on a very detailed analysis of specific types of cells in terms of their internal variability. Until recently, the problem of identifying cell subpopulations in a biological data set was based only on the manual interpretation and assignment of cells due to their phenotype or morphology [26]. Nowadays, due to the increasing automation of biological data acquisition processes and the introduction of more and more accurate methods of obtaining these data, such as the single-cell technology generating particularly high-dimensional data, it is also very desirable to automate the processes of recognizing specific cell structures, or general analysis of biological data sets. Such automation aims to accelerate the recognition of new, unknown, and rarely occurring cell types. The immune system is an excellent field for such analyzes. Given the high diversity of all pathogens and having high adaptive capacity, the immune system, by definition, must contain an enormous variety of individual cells differing in type, subtype, or phase of the cell cycle. With such advanced technology as single-cell sequencing, the only limitation is the availability of algorithms enabling an insight into the detail of the composition of biological matter and the processes taking place in it. Because of this technology's introduction, it became possible to look even deeper into specific types of cells, thus dividing cells previously assigned to the same category into smaller classes [27] [28]. Considering the discovery of white blood cells in 1843, any further technological advances, whether in obtaining biological data or complex analysis, made another building block leading to the present state of knowledge about the functioning and heterogeneity of these blood components. Moreover, it has also become possible to better understand the functioning of individual cell subtypes in the context of pathways of biological processes. This has made it possible to understand many previously unclear aspects of the immune system’s functions and adaptations. It is thanks to the simultaneous development of ever more accurate and faster computer analysis and data visualization technologies, as well as the accompanying enormity of acquired data that scientists have had the opportunity to view increasingly complex structures, explaining the heterogeneity of biological data at a lower and lower level of complexity [29]. This is why there is a growing need to develop more automated techniques. Using complex data analysis protocols based on manual foundations is sometimes even impossible. Therefore, there is an increasing need to produce universal automatic tools for processing and in-depth data analysis with the lowest possible level of human interference.

Despite being divided into separate subpopulations showing differences in detailed functioning and participation in specific biological processes, white blood cells have some common features. These cells arise in the bone marrow and can be found successfully in blood and lymph tissues. They are a crucial part of the immune system, to which we owe the ability to maintain life in contact with various pathogens. Moreover, looking globally, the broadly understood evolution phenomenon can occur. White blood cells fall into two main groups: granulocytes and agranulocytes. Instead, these two groups fall back into many more minor subtypes (depending on how accurately we want to judge their

division). Granulocytes are broadly divided into neutrophils, eosinophils, and basophils. Conversely, agranulocytes contain subtypes such as lymphocytes (including T-cells, B-cells, and NK-cells) and monocytes. However, the most significant percentage is due to neutrophils, which constitute approximately 50-70% of WBCs. Subsequently, due to the percentage share in the WBCs fraction, there are lymphocytes (25-35%), monocytes (4-6%), eosinophils (1-3%), and basophils (up to 1%) [30]. T-cells constitute the most numerous fraction among lymphocytes, occupying 80 to 90% of its fraction, leaving B-cells second and NK-cells third in the percentage share.

Both T-cells and B-cells are responsible for fighting diseases. Their spectrum of activity, however, differs slightly [31]. T-cells respond to viral infections, support other cells' immune function, and destroy cancer cells. On the other hand, B-cells are specific in response to bacterial infections. Monocytes, by transforming into macrophages, gain the ability to eat microorganisms and get rid of dead cells. This increases strength and affects the proper functioning of the immune system. The most abundant neutrophils allow microbes to be trapped in infections by ingesting or destroying them with enzymes. Besides, basophils and eosinophils produce necessary enzymes in allergic reactions, inflammations, and parasitic infections.

A fascinating study published by R. C. Wilkins et al. [32] in 2002 showed that white blood cells have significantly different cellular responses to the radiation factor. This analysis considered subpopulations of granulocytes, B-cells, NK-cells, and T-cells (divided into two subgroups). Utilizing a modified neutral comet assay, the apoptotic fraction of the control sample and the sample irradiated with X-rays were analyzed. The experiment was carried out in an in vitro environment. As a result of examining the apoptotic fraction of control cells, it was shown that the highest fraction of spontaneous apoptosis characterized granulocytes. The subpopulation of B-cells and NK-cells also showed a high index, while the lowest index described the T-cells fraction. Moreover, as a result of cell irradiation, the granulocytes again showed the highest value of the apoptotic fraction, while the most significant increase occurred among the T-cells subpopulation. This indicates a very high sensitivity of the T-cells subpopulation to ionizing radiation.

2.6 Classification problem

The problem of cell classification is not trivial, mainly due to the complex structure of biological data. Their increasing volume, both in the context of the classification entities and the number of features, is caused by developing subsequent, more detailed data acquisition techniques. Two concepts are fundamental in the case of the classification of high-dimensional data. The first is data disruption that often occurs with modern, meticulous data acquisition techniques. Due to their high accuracy, the importance of controlling the disturbances is also increasing so that the analytical procedures are entirely focused on recognizing the heterogeneity of the sets or analyzing the factors specified for testing. The disorders arising at the data preparation stage may effectively hinder or prevent identifying specific data differentiating factors. The second essential aspect related directly to the high dimensionality of the data is the increase in the number of features irrelevant from the point of view of the research problem. Analyzing all generated and available features often leads to a significant and unprofitable increase in computational and physical costs for analyzing such complete data.

The classification problem is a two-step process without considering the apparent differences between the existing methods. The first stage is based on constructing the model. It consists of the learning and validating processes. Separating the train, validation, and test sets cautiously is essential. They should reflect the analyzed data as best as possible, without introducing additional disturbances in the critical process of model learning, in the form of, for example, uneven or unreal distribution of cells in subsets. The train and validation sets are used to build the model. The train set is utilized to develop appropriate weights for the introduced features. Then the validation set is iteratively used to self-correct the estimated weights or to control the learning process. The second step of building a fully functioning model is based on fine-tuning the ability and quality of classification of individual

observations. Instead of classification, the goal is to create a complete profile of the differentiating features of a set of observations.

Currently, there are many different classifier-building methods. One of them is the quite widespread method of decision trees. The undoubted advantage of this method is the relatively quick construction of the entire decision tree compared to other methods. This method makes it possible to determine whether an element belongs to a specific class based on a series of decisions [33]. Moreover, this method can be successfully used in the case of a multi-class problem. The deeper we look into the resulting decision tree, the more complicated the decision and the rules for determining the affiliation of the observation. On the other hand, decision trees have one major drawback; it is only possible to discover the correlation between the data with the need to develop additional calculations.

Another exciting example is the neural networks method. These are computational systems inspired by the operation and construction of biological neural networks found in the brains of animals. Collections of interconnected nodes are called neurons, which are supposed to reflect neurons in biological brains. Neurons communicate through connections that carry information. Each neuron performs appropriate calculations based on applying nonlinear functions to each input signal and transmits the result to the subsequent neurons connected. A crucial element in constructing the entire network are weights assigned to signals from individual neurons [34]. Such signals can increase or even decrease the strength of the conducted signal. Moreover, neurons are often grouped into layers between which signals can be transformed differently [35].

The following solution for classifiers inspired by natural biological processes is the genetic algorithm, belonging to a larger group of evolutionary algorithms [35]. This method is based on the elements of Darwin's theory of evolution based on aspects such as mutation, selection, and crossover. These algorithms are based on determining the match of a specific element of the data set compared to the other components. These procedures are quite complicated in computation and require multi-stage action, but they can solve very complex problems, often requiring a very long time to solve.

Another example of the algorithm successfully used in the classification problem is the k-nn nearest neighbors method. To simplify, this algorithm predicts the membership of a given observation based on the set of the n-nearest neighbors of this observation. What is essential is that the affiliation of neighborhood observations is known [36]. The closest neighbors are usually defined using the minimization of a simple distance metric such as the Euclidean metric. The predicted value is based on averaging the values of selected observations. This method is beneficial in the case of data analysis with a complex relationship that is difficult to model using other methods.

The last mentioned in this dissertation approach is a group of statistical methods that use more or less complex mathematical expressions for modeling. One of the subsets of statistical methods is logistic regression, and it can be successfully used when there are only two predictive groups: positive and negative (binary classification). This method determines the probability of belonging to a positive class of a data set element based on the calculated and numerically presented relationship between the independent and dependent variables. The expected observation value is the sum of the products derived from the observation value and the coefficient fitted to the independent variable [37]. The resulting coefficients of the independent variables determine how they affect the dependent variable. The element is assigned to one of the two classes based on the estimated probability value for the specified observation.

Such a large variety of classifier construction methods is not only reflected in the fast development of high-dimensional data analysis techniques. An even more important reason is the large variety of data that can be subjected to the problem of classification and, above all, the assumptions and challenges of the research problem. To a large extent, the choice of the appropriate method is dictated by the aim of the analysis because it imposes on us a particular research thesis assuming differences or their absence from the analyzed phenomenon. Very often, it may turn out that the time-consuming and complicated analysis methodology not only takes a disproportionate amount of time concerning the expected results but may cause overfitting or overtraining of the classification tool. The result of such an action is a joint waste of time and a result that will be useless in achieving the set goal. It all means

that choosing the most computationally complex method does not always produce the best results. In this case, particular attention should be paid to the costs and benefits of the analysis performed. A fundamental question should be answered: will the results obtained by a much more complicated and labor-intensive method of constructing a classifier allow for significantly more accurate results? An insightful and critical answer to this question will enable us to avoid many failures and disappointments and often speed up the analysis period from obtaining data to achieving the set goal.

2.7 Feature selection methods – a literature overview

The feature selection problem is fundamental in building an efficient and accurate classifier. In the context of the available features, the dimensionality of the data set undoubtedly impacts the computational and time costs necessary to achieve this goal. Therefore, dimensionality reduction is a desirable step before the model-building approach. This allows the number of features to be limited to those that are important from the point of view of the research problem under consideration. The phenomenon of feature selection has been described in detail in the manuscript entitled *Feature selection methods for classification purposes* [38].

There are two types of possible operations on features to reduce dimensionality. The first is feature extraction, which usually reduces dimensionality significantly, but completely new features are created, resulting from a combination of input features. The very high informativeness of the set of features is maintained, with the possibility of a significant dimensionality reduction. What is important, the features created in this way, and analyzed in further stages, are not interpretable from the biological point of view. This is due to the creation of new features resulting from the combination of input features without the possibility of disconnecting them again. Principal Components Analysis (PCA) is a widely used feature extraction example. On the other hand, the second dimensionality reduction method is called feature selection. It consists of selecting statistically significant characteristics, for example, differentiating two data samples, without transforming the input data. This way, the remaining features' full biological interpretability is preserved. Because the task is closely related to the subsequent interpretation or manipulation of selected features in large part of the research work, the main types of the feature selection procedure are described below.

In general, feature selection methods include unsupervised and supervised approaches. Unsupervised feature selection refers to the process which does not need the output label class for feature selection, and this type of approach can be used successfully for unlabelled data. On the other hand, supervised feature selection refers to the method which uses the output label class. It uses the target variables to identify features that can increase the model's efficiency. Supervised methods are divided into the wrapper, filter, and hybrid methods [39].

Filters are not very time-consuming because they do not use complicated machine-learning algorithms [39]. Without extreme computing resources and a significant amount of time, the data dimensionality can be reduced successfully. It is easy to implement most filter methods because they are mainly based on the statistics that enable the creation of feature ranks [40]. Features are dropped based on their relation to the output or how they correlate to the output. The critical point when using the method from the filters group is the selection of an appropriate cut-off threshold for the number of selected features [41] [42]. This task is challenging because specific decisions must be made about which features to retain in the analysis and which do not contribute the appropriate and expected level of informativeness based on the chosen and estimated characteristics. After this, a list of features is obtained that, when applied to the dataset, will reduce the data dimensionality. Filters are quite a general approach, so they are especially recommended for high-dimensional data [43]. Another advantage of filters is that they are also easy to interpret – a feature is discarded if it has no statistical relationship to the target variable. On the other hand, however, filter methods have one major drawback. They look at each feature in isolation, evaluating its relation to the target. This makes them prone to discarding valuable features that are weak predictors of the target but add much value to the model when combined with other features [40].

The next type of supervised method is the wrappers approach. This method uses a learning algorithm to select the most critical, differentiating features. However, the use of a complex methodology requires substantial computational resources, especially in the case of high-dimensional data [43]. Therefore it is not a thoroughly recommended method for complex datasets [44]. The wrapper algorithm splits the input data into subsets and trains a model [39], which is then used to score different subsets of features to select the best one. There are very often utilized backward and forward selection procedures. Backward selection is an approach in which we start with a full model containing all available features. In subsequent iterations, we remove one feature at a time. The following most common procedure is forward selection. It works in the opposite direction compared to backward procedures: we start from a null model with zero features and add them one at a time to maximize the model's performance. Building a set of features ends when the stopping criterion is met in both cases. It can be, for example, a difference in the value of the quality metric obtained for the compared sets of features or a deterioration in quality for a specific set of features.

There are two key aspects to compare the two supervised feature selection methods. Both approaches have many undoubted advantages but are also characterized by disadvantages that should be remembered when considering a specific research problem. One such aspect is the time necessary for the feature selection procedure and the computational cost. In this regard, wrappers are much more demanding methods than filters [40]. Wrappers are computationally expensive and very time-consuming procedures, especially in the case of high-dimensional data [42]. On the other hand, filter methods use statistical measures to evaluate a subset of features, making them less demanding in terms of computational cost. But apart from the discussed aspects, a fundamental problem is the purpose of feature selection and what losses and disadvantages of a given approach we can accept. For this reason, it is worth remembering that filter methods measure the significance of features by their correlation with the dependent variable, while wrapper methods measure the usefulness of a subset of features by actually training a model. It also means wrappers can catch correlations between subsequent features [45]. Suppose we care about maintaining the dependencies between individual features, an essential aspect of the problem under consideration in which we must keep the correlation between features. In that case, this should be a solid guiding argument in the considerations; thus, the focus should be paid to wrappers that give such opportunities. Filters cannot capture such relationships because they examine the impact of individual features on the dependent variable, considering them entirely separately.

As mentioned earlier, the hybrid approach is a third type of supervised feature selection method. It combines both wrappers and filters into a single system. It gives the most significant scope for manipulation and selection regarding the intended goals. The use of filters, reducing the dimensionality of the data, significantly affects the reduction of computational resources and the time necessary to perform complex methods included in the category of wrappers. Therefore, adequately selected techniques from the filter and wrapper categories, forming a hybrid approach, can eliminate many disadvantages of these approaches used separately [46]. However, especially in the case of hybrid methods, attention should be paid to the weaknesses of individual systems. Improper and ill-considered application of the random order of occurrence of techniques from the category of filters or wrappers may result in incorrect and misleading results [47]. Using the filter method first, we can filter out some significant features because this approach looks at each feature in isolation. It does not consider if a specified feature can correlate with other features, making it quite crucial from the point of view of the analyzed problem. The wrapper method controls possible correlations between all the features, so no information is lost in this context. Due to the possibility of excluding some correlation-important features using the first filter and then wrapper approach, the resulting model can be weaker regarding the model's classification quality compared to the opposite approach, where the model is as strong as possible for a given set of features. However, using filters first and then on the reduced dataset utilizing wrappers can significantly speed up calculations and reduce the problem of data complexity. More favorable calculation times can be achieved by reducing the number of features. Choosing the proper method from filter and wrapper categories is problematic in the feature selection approach. The biggest and much more time-consuming problem is to select the correct workflow for feature selection and

consider all the advantages and disadvantages that we can accept in the context of the analyzed research problem.

An essential question may also be asked – why is feature selection such a key and important step in high-dimensional data analysis? The first answer that comes to mind is overfitting. Suppose there is a relatively large number of features to the observations. In that case, the model can quickly match the target function on the training data. Such an approach will not result in a generalized model, which is often the goal. Dealing with big data is always connected with variables that, in part or even in the majority, do not carry any significant information from the point of view of the analyzed problem. This means they have no relation with the target and are entirely unrelated to the task the model is designed to solve. Filtering out irrelevant features will prevent the model from picking up on false correlations it might carry. Sometimes created models suffer from what is known as the curse of dimensionality. It means that in a very high-dimensional space, each training example is so far from all the other examples that the model cannot learn any valuable patterns. The solution is to decrease the dimensionality of the feature space to prevent this phenomenon. Additionally, with too many features, we also lose the explainability of the model. While interpreting and explaining the model's results, it is essential to remember that the more features, the more complicated it will be to find key aspects that drive the analyzed problem. The last aspect is the phenomenon called Occam's Razor. It means that simpler models should be preferred over the more complex ones as long as their performance is the same or not significantly different in terms of the intended goal. This shows that a properly carried out feature selection makes it possible to solve many classification problems, select a small set of features relevant to the problem being solved, significantly reduce the dimensionality of the data, and allow to make a clear explanation of the analyzed problem.

Moreover, feature selection is critical if we know that many research problems, especially those with a biological basis, must be reflected in the researched field of science. It is worth remembering that not always the solutions proposed by the majority can be applied to the analysis of our data. We must think carefully and plan the entire feature selection path concerning the expected results. The use of universal solutions that give benefits in the form of time savings and reduction of computational costs carries a specific price. The question is whether this price is acceptable in the context of our work. Sometimes the gains from retaining more information are so small that they can be neglected and thus reduce computation time. However, in some cases, if even the smallest piece of information counts and there is the possibility of significant correlations between features occurrence, it is worth not taking shortcuts.

3 Materials

In this dissertation, there are analyzed two high-dimensional datasets. Both sets are derived from single-cell sequencing experiments based on white blood cells and are technical repetitions of the same trial. To maintain clarity and legibility of references to individual data sets, they have been presented and described as sets A and B. Both datasets include two cell samples: normal (control) and ex vivo irradiated. The irradiated group of cells was exposed to ionizing radiation at a dose of 1 gray (Gy), which is the energy of 1 joule (J) absorbed by 1 kilogram (kg) of the irradiated environment. The analysis of complex high-dimensional data in the form of two technical repetitions of the same experiment not only allows for achieving the individual goals of this work but also gives a clear insight into possible errors or inaccuracies related to the procedure of the conducted experiment. Moreover, it enables the creation of designs resistant to potential differences in sets from technical repetitions of the same experiment. Based on the analysis of both data sets, it is also possible to capture the accuracy and quality of the procedures for preparing and collecting the analyzed data.

3.1 Single-cell RNA-sequencing data

As mentioned, the analyzed ex vivo data consisted of control and irradiated samples (24h post-irradiation time point). Both samples were obtained based on the white blood cell population. The blood control sample, after being collected, was stored at 37ºC. A part of the control sample was subjected to ionizing radiation at a dose of 1 Gy. To keep both control and irradiated samples at -80ºC until RNA extraction, the RNAlater was added to both samples. In the case of the irradiated sample, the RNA later was added 24 hours after the occurrence of the radiation agent.

The data acquisition and processing of single-cell sequencing was performed according to the protocols of the BD™ Single-Cell Multiplexing Kit. This platform uses a cartridge workflow and a complex barcoding system. Control and irradiated cells are marked with a specific SampleTag to be able to process them together in subsequent steps and allow them to be separated later when the appropriate procedures are completed. Such prepared control and irradiated cells are loaded onto a matrix of microwells, where they fall into weels with gravity. Next, the mixture containing beads is added to the cell compound. Then beads are deposited in the wells containing subsequent cells. The microwells are designed so that only one bead can fit in there. When the lysis process starts, RNA is released from the cells and bound to nearby beads. A magnet pulls the biological material together with the bound beads. The beads' design allows binding to only one cell. Beads from the BD Rhapsody platform have a two-step control ensured by the barcoding system. It is possible to recognize the cell (poly-dT primers) from which the RNA is derived and the specific transcript (UMI) of origin. In the last step, a PCR procedure is carried out in which cDNA is synthesized using reverse transcriptase. The entire process guarantees the formation of a labeled transcriptome of thousands of introduced cells [48]. Before sequencing, quality control is also performed using Agilent Bioanalyzer with the High Sensitivity Kit and the Qubit dsDNA HS Kit. In the analyzed data, the Illumina® paired-end sequencing platform was used for sequencing purposes.

The BD Rhapsody™ platform also provides the processing of raw sequencing data into the count matrices necessary for bioinformatics analysis. The tool's algorithm work with raw paired-end sequencing data from Illumina sequencers. In simple terms, it is possible to carry out the following analysis steps: filtering by the quality of reads, annotation of R1 and R2 reads, and combining information from both read annotations and molecules. The last mentioned step consists mainly of removing possible PCR reactions and sequencing errors, using the implemented RSEC and DBEC algorithms based on UMI determinations. The next step in quality control is the removal of erroneously generated cell labels that may have arisen in preparing data for sequencing. The last step necessary to obtain the count matrix is sample multiplexing. This is due to loading multiple labeled samples (control

and irradiated) into the BD Rhapsody Cartridge. This way, all samples can be analyzed simultaneously, maintaining the same experimental conditions. After the data collection and preparation procedures are completed, cells belonging to individual samples can be easily separated and explored in later steps according to the research thesis needs.

3.2 Count matrices

The data analysis workflow in this doctoral dissertation was adapted to the count matrices form created by the BD Rhapsody™ system. The rows in the generated count matrices are represented by the list of cell indices entered for analysis. The columns contain the list of genes of the used immune response panel. Due to such a matrix design, each cell in the matrix structure represents the number of molecules detected in each cell per specified gene. Moreover, the raw sequencing data analysis platform also generates a file containing information about the quality control results performed for individual genes. This file marks each panel gene with one of the three statuses: *not detected*, *low depth*, and *pass*. Status *not detected* indicates that the gene was present in the panel but not in the real data due to its zero counts. *Low depth* status means the minimum required sequencing depth has not been reached. The only status indicator that allows for further analysis is the *pass* status. It indicates regularities in the counts and sequencing depths for a given gene. Moreover, when the multiplexing option is selected, what is required while analyzing multiple samples simultaneously, an additional file is generated. It includes the marks necessary to assign individual cells to appropriate origin samples. This file contains the index of each cell and its associated SampleTag.

3.3 Data summary

Table 2 summarizes the cell per gene count information in the two ex vivo technical replicate datasets (set A and B).

Table 2. The number of cells and genes in the subsequent ex vivo datasets.

Dataset	Control cells	Irradiated cells	Other cells	Total cells	Total genes
Set A	1584	1139	1516	4239	452
Set B	2301	1988	2633	6922	452

The first two columns, described in the table, show the number of control and irradiated with 1 Gy dose cells. The third column, pointing to *Other cells*, represents the cells identified by the filters applied by the BD Rhapsody platform that did not pass the initial quality control steps. Among the designations used, indicating the reason for cell rejection, there are *Mixed*, *Multiplet*, and *Undetermined* cells. *Mixed* group cells are cells without an assigned control or irradiated SampleTag. The other two statuses, *Multiplet* and *Undetermined*, indicate cells that could not be correctly classified during the experimental process due to technical issues. These three groups of cells (*Mixed*, *Multiplet*, and *Undetermined*) cannot participate in further analysis methods, such as irradiated cells' genetic profile detection, classification, or even cell subpopulation recognition. Therefore, the analytical procedures adopted in the dissertation were applied only for two correctly marked and processed groups of cells, i.e., for cells from the control and irradiated samples.

4 Publicly available tools and developed workflows

The doctoral dissertation utilized publicly available free tools and own-developed workflows using several statistical and machine learning methods. Publicly available tools such as UMAP and HDBSCAN were used to visualize datasets with unsupervised learning methods and recognize individual white blood cell subpopulations. Own-developed workflows were used in a wide range, including the main aspects such as feature selection, building models based on logistic regression and neural network methods, classification of control and irradiated cells, and building a genetic profile of cells irradiated with a 1 Gy dose. In the following subsections, there are described both the methods and tools used.

4.1 Publicly available tools

4.1.1 UMAP

Uniform Manifold Approximation and Projection (UMAP) [49] is a free and publicly available tool. It is beneficial in the context of high-dimensional data visualization. The high complexity of the analyzed data is characterized by the speed of operation at an acceptable level. This tool is highly effective in reproducing the global structure of the entered data. Moreover, it often enables the detection of hidden structures inside the datasets, depending on the input data specificity.

The instructions for constructing a weighted graph are implemented and utilized in the initial stages of the algorithms' operation. Inside the graph, the marginal weights are related to the probability of connecting two specific points, which depends on the proximity of these points to each other. In simple terms, the weight of the edge is higher the closer the following points are. The possibility of connecting points depends on the neighborhood's radius because it is possible to connect points only within the vicinity. The radius of the vicinity of a given point is determined based on the distance to the *n*th nearest neighbor. When the designated circles with a certain radius for two points are superimposed to a certain extent, it is possible to connect them with a weighing edge. Delving into the process of creating a high-dimensional structure, it is worth mentioning the Čech complex, which is precisely a method of combinatorial topology representation with the use of sets [50]. This creates a high-dimensional graphical structure optimized for a low-dimensional form. Such optimization works by creating a low-dimensional graph that is as close to a high-dimensional graph as possible, taking advantage of the insight from Riemannian geometry and algebraic topology. The basic unit in the optimization phenomenon is simplex, a *k*-dimensional object created from the combination of k+1 points. By joining points with overlapping radii, it is possible to create more complex, multidimensional simplices. Going further in this theory and considering the data set as a set of simplices, it is possible to capture a topology representation. It turns out that most of the topology mapping focuses on the 0- and 1-simplices that make up the Vietoris-Rips Complex [50]. By considering only 0- and 1-simplices, it is possible to project a high-dimensional topology onto a low-dimensional topology. Moreover, this approach significantly reduces the computational costs and thus the directly related time-consuming process. The most complicated step is to select the appropriate radius to define the number of *k*-nearest neighbors. Choosing a too small radius will result in the formation of small, local clusters. On the other hand, selecting a too-large radius, the included instructions will lead to connecting all available points. Hence, UMAP uses an approach that does not directly define a radius value but uses a variable radius. This value is predefined using the *k*-nearest neighbors. Thus, UMAP sets a radius equal to the distance to the *k*th nearest neighbor. It is essential in the case of high-dimensional data where the points are closer and closer to each other with the increase in dimensionality. The fuzzy connections between the points are then created within the radius. The connection weight value is determined based on the distance between neighboring points. However, there is a problem with edges connecting two points with

different values of connection probability, of course, directed in opposite ways. To determine the weight of the joining edge, the UMAP tool calculates the probability that at least one of the edges exists.

When applying the UMAP tool to real data, selecting the appropriate parameter values [51] is essential. The most critical parameter, *n_neighbors*, is the one responsible for the choice of the *k*-nearest neighbors. Thanks to this parameter, it is possible to control the local and global structure of the data. As previously mentioned, the low values of this parameter will focus on the local structure. In contrast, high values will allow insight into the global structure at the cost of losing some detailed information. The second crucial parameter is named *min_dist*. It describes the minimum distance between points in the low-dimensional space. Therefore, this parameter determines the point concentration. For this reason, using high values of this parameter will result in a better reflection of the global data structure due to the smaller data grouping.

The UMAP tool allows working in both a supervised and unsupervised manner. This is undoubtedly a great advantage of this tool. It can be widely used for various purposes, often having a joint part in one test procedure. Working with this tool undoubtedly reduces the time necessary to perform complex analyzes that are very computationally demanding. It allows, among other things, to discover the local and global structure of the analyzed data, capture unexpected behavior of a data set observations, or contribute to a deeper analysis of dependencies occurring in a data set. Additionally, it is possible to perform supervised learning for later classification observations with an unknown origin.

4.1.2 HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [52]is a clustering algorithm that uses hierarchical clustering techniques and the phenomenon of cluster stability. Its operation generally consists of five basic steps [53]: density estimation, construction of a minimum spanning tree based on a distance-weighted graph, hierarchization of connected clusters, condensation of clusters, and selection of stable clusters from a condensed tree.

The first step in density estimation is crucial to understand the underpinnings of this algorithm and draw attention to the differences from other clustering algorithms. Density estimation is used to generate a map that shows the clusters of points in the analyzed data. On the other hand, in the case of real data, which is often characterized by a high content of noise and multi-cause disturbances, single points may cause the joining of clusters that are predisposed to separate and create spatially separated structures. For this reason, lowering the noise density level is necessary by introducing the mutual reachability distance metric [53]. Of course, this metric uses the primary *k*th nearest neighbor metric and the core distance metric [53]. The core distance metric described the circle's radius drawn for the specified center point containing the *k* nearest points and established for the *k*+1 point touched by the circle. The lower the value of the radius of the circle defining the core measure, the more densely the point is in the region. Based on the mutual distance between two points and the core measures for both points, the value of mutual reachability distance is determined as the highest value by (1) [53].

$$d_{mreach-k}(a,b) = \max\{core_k(a), core_k(b), d(a,b)\} \quad (1)$$

The next stage, i.e., constructing a minimal spanning tree, is based on the previously calculated metric. The weights of the edges connecting the points are equivalent to the mutual reachability distance metric values. Building a minimal spanning tree based on Prim's algorithm was used. Edges are added to the tree one by one, selecting in a given step the edge with the lowest weight connecting the tree to be built with the vertex, which has yet to be included in the tree. Based on the generated tree, the next step is to convert it to make a hierarchy of connected clusters. Joined clusters of points are created, starting from the edges with the lowest weight values and working toward the higher values. Building a cluster hierarchy ends when one central point cluster is reached. On the other hand, the problem of choosing the optimal number of clusters remains unsolved. Of course, it is possible to enter one tree

cut value, but in addition to the clusters through which the cut line passes, we also get a cloud of single points (below the set cut value) that must be declared as noise. In this case, there is a perfect chance to connect individual clusters or classify many points as noise. Hence, in the HDBSCAN tool, there is a cluster extraction solution based on the condensation of the cluster tree. In assessing what a cluster is, it turned out that it was necessary to introduce a parameter defining the minimum number of points that would be perceived as a separate set and not counted as noise. In the condensation of the tree, it is essential to re-traverse the established cluster hierarchy at each stage of the cluster division to determine whether the newly created cluster meets the minimum cluster size requirements. If so, another division of the home cluster is created. On the other hand, if these requirements are not met, the points that would make the next split are classified as "dropping out" "points from the home cluster. These points are then marked as noise. As a result, we get a more condensed picture of the hierarchy and division of clusters with a much smaller number of nodes.

Cluster extraction is the last step required to achieve the intended purpose of splitting the dataset into clusters. To decide on the number of selected clusters, two lambda values were introduced to define the duration of the cluster. The first is the cluster birth value, which begins with the division of the home cluster and the formation of the new structure. The second is the death of the cluster (which intuitively does not apply to every distinguished structure) determined by dividing the cluster into smaller structures. The difference between the value of cluster death and the value of cluster birth is determined by the cluster stability measure, which can be called the cluster lifetime. It is worth emphasizing that the death of a cluster is also considered in the case of "falling out" of points defined as noise (the tree condensation procedure described earlier). So how are clusters of final choice determined? This is where HDBSCAN computes the child clusters' stability sums starting at the tree's bottom. If this sum is lower than the stability of the parent cluster, then the selected cluster is the parent cluster, and the following divisions are meaningless. After reaching the root, the tool returns the clusters chosen as the optimal version of the dataset split.

However, we often want to assign each analyzed point to the closest cluster in many real data analysis approaches. The standard clustering approach assigns points arbitrarily to individual clusters or noise. The solution to this problem is introduced in the HDBSCAN tool extension called "soft clustering" [54]. In this solution, points are not assigned directly to the clusters, but the probability vector of their assignment to each generated cluster is determined. This allows the algorithm to choose the point’s affiliation more accurately and have insight into mixed affiliation. For real data analysis, and especially when analyzing biological data, this can provide additional insight into the behavior and internal structure of the data. It is also possible to analyze the strength of the point binding to the clusters to which specific biological functions will be assigned by further analysis. This approach gives unique possibilities for research and the case of a very detailed study of the internal structure and different phenomena occurring in the data set. The soft clustering method is based on introducing two metrics that define the distance of a point from the clusters. One of them is the usual distance from individual clusters. The point deviations from the separated clusters determine the second. However, calculating the distance of a point from individual clusters is not a trivial task, mainly due to the frequently occurring uneven shapes of structures. For this reason, the set of exemplary points that defines the cluster’s center is used to assess the distance [54]. As a set of exemplar points, observations are presented that last the longest in a given cluster, i.e., have the most extended lifetime. The second measure, as was mentioned, is related to determining how strongly the point deviates from the designated clusters. To estimate the value of this measure, the membership length of a given point is compared with the total duration of a given cluster. The combination of these two metrics is created using Bayesian conditional probability. As a result of the generally described "soft clustering" procedures, it is possible to have an insight into the probabilities of individual points belonging to selected and discovered clusters.

It is worth mentioning that the HDBSCAN utility allows the user to change several parameters. Of course, the most important are parameters called *min_samples* and *min_cluster_size*. The appropriate selection of these parameters may be of crucial importance in the results of the work utilizing this tool.

A reasonable approach is the introduction of a proper metric informing how the selected set of startup parameters affects the separation of clusters. Of course, the metric and the inference should be labeled primarily in conjunction with the expected results and the purpose of the analysis. Inattentive juggling with startup parameters can lead either to the loss of essential divisions by treating the data set too general or to dividing into clusters that do not show statistically significant differences in terms of the stated purpose of the analysis.

4.2 Developed workflows

The reason for developing a new workflow related to analyzing data from single-cell sequencing experiments was primarily the lack of a developed and publicly available method for feature selection and building a model based on the genetic profile of irradiated cells. The available processing methods require many tools in which it is impossible to ensure complete control of the flow and analysis of the entered data. The feature selection stage is a crucial moment in this work. Indeed, lack of control at some stage will result in inaccuracies and distorted results in the form of a genetic profile of irradiated cells. As a consequence, the cell classification process will also be severely affected.

4.2.1 Logistic regression-based workflow

The algorithm based on machine learning (ML) methodology, implemented as part of the doctoral dissertation, aims to generate a model describing the gene signature of irradiated cells. In the case of the research problem, binary logistic regression was used, and which task was to learn the classifier to assign observations to one of two classes: positive (irradiated) or negative (control). Logistic regression estimates the vector of weights assigned to the model's components. These weights are information about the importance of a given feature in the problem of class recognition. The weights can take both positive values, showing the impact on the classification of a positive class, and negative ones, showing the effect on the classification of a negative class [55]. To calculate the value of the probability of a particular observation belonging to a positive class, it is necessary to use the sigmoid function. The decision on belonging to a positive class is made using a simple decision procedure – the observation is classified as being in a positive class if its probability of a belonging value is at least 0.50. Learning weights, i.e., model parameters takes place in a supervised manner. This means that the true class for a particular observation is known. The model aims to make class predictions for that observation as close to the real class as possible. The loss function determines how much the predicted class differs from the actual class. In the proposed approach, the inverse of the loss function was used due to the subsequent need to maximize it when determining the optimal model using the Bayesian Information Criterion (BIC) value. The model parameters are calculated iteratively, striving for the best fit by minimizing the differences between subsequent likelihood values. Below is a thorough and detailed description of the implemented approach and the formulas used.

The implemented ML algorithm accepts three arguments as input values: *LR* - training set matrix (genes/features in columns, cells indices in rows manner), *anno* - sorted, as in the input matrix, a single-column matrix containing the markings of a cell belonging to one of the two samples, *draw* - number analyzed genes (equal to the number of columns of the transferred *LR* matrix). As part of the implementation, changing four basic parameters is possible. Parameters influencing the functioning of the algorithm and the results achieved are *epsilon* - acceptable likelihood difference in the process of learning parameter values, *max_iter* - maximum number of iterations of parameter value estimation (stop criterion), *alpha* - learning rate (default value is set to 0.001). An additional parameter influencing the readability of the obtained results when iteratively invoking the algorithm several times is called a *draw*, which means the run number (total model number). At the output of the implemented model, there is information saved in .txt files containing details about the models generated in each iteration (*draw*) regarding Bayesian Information Criterion values, Log-Likelihood (*LL*) values, weighted classification quality values based on the validation set (*Acc*), the names of the genes included in the model in the order of their appearance in the final model (*gene_step*), and the values of the final model

parameters in the order in which the genes appear in the model (*parameters*). The primary function that is the basis for calling the algorithm (*FullLRclassifier)* is responsible for all the procedures for features selection utilizing the wrappers method and building the final model. Inside the primary call function is the *ClassifierLR* function, which is responsible for the parameter values learning process for the given set of genes involved in modeling.

The overall operation procedure of the program, to simplify the complex structure, can be divided into building single-factor models and building multi-factor model parts. Both methods follow each other after running the algorithm. The process of building models begins with calling the function along with the passed variables and parameters. After the primary function accepts the input values, building one-factor models follows (the complexity of the created models in the context of the included features is N=1). Subsequently, as part of the transfer of a single gene to the second function, the probability of the cells belonging to the positive class y=1 given a set of genes *x* and the collection of θ parameters is calculated (2), and their labels are predicted using the sigmoid function (3) and decision scheme.

$$probability_{cell} = \frac{1}{1 + e^{-(\theta_0 + \sum_{gene}^{N} \theta_{gene} * x_{gene})}} = P(y = 1|x;\theta) \quad (2)$$

$$prediction_{cell} = \frac{1}{1 + e^{-z}} \quad (3)$$

Where:

$$z = \sum_{i=1}^{length(\theta)-1} \theta_{i+1} \times LR_{x,i} \quad (4)$$

This function gives results in the form of the probability that an observation belongs to a positive class. Using a simple decision scheme, cells are classified into the positive (1) or negative (0) group.

if	$prediction_{cell} < 0.50$	then	$\widehat{group}_{cell} = 0$
else		then	$\widehat{group}_{cell} = 1$

Model parameter values are calculated using gradient descent (5) with a learning rate α = 0.001. Finding the optimal set of parameters for the analyzed model is possible using the descent gradient.

$$\theta_{cell}^{new} = \theta_{cell} - \alpha \sum_{cell=}^{Nc} \left[(probability_{cell} - group_{cell}) * x_{cell}^{gene} \right] \quad (5)$$

Therefore, the model learns the parameter values in such a way as to achieve the best results in the context of the selected loss function, which for logical reasons, of applications in this implementation takes the form of the inverse of the loss function (6).

$$LL = \sum_{CellsPositive} \ln\left(probability_{cellPositive}\right) + \sum_{CellsNegative} ln(1 - probability_{cellNegative}) \quad (6)$$

At this stage, the likelihood difference (7) value is also checked for subsequent calculation iterations for the values of the model parameters. If this difference is smaller than specified by parameter *epsilon*=0.1, the procedure of calculating parameter values is completed for a given model, and these parameters are considered final.

$$L_{dif} = |L_i - L_{i-1}| \quad (7)$$

Where:
i is the parameters estimation iteration
and L is the likelihood estimated based on the formula (8)

$$L = \sum_{CellsPositive} probability_{cellPositive} + \sum_{CellsNegative} (1 - probability_{cellNegative}) \quad (8)$$

The procedure for a single-gene model ends when the BIC (9) and LL (6) values are calculated, and the declared maximum number of iterations or L_{dif} is reached.

$$BIC = N_{parameters} \times \ln(N_{cells}) - 2 \times LL \quad (9)$$

The calculations for the single-feature models with the following transferred genes begin. When the analyses are performed for all single-factor models based on genes available in the shared data matrix, the BIC values for all these models are compared. The model with the lowest BIC value is selected best from single-factor models. For this purpose, the inverse of the loss function was maximized, which results directly from equation (9) and the presented algorithm workflow. After selecting the best single-feature model, the procedure with the degree of complexity N=1 ends.

Building multivariate models start at the level of complexity N=2. The remaining genes not included in the univariate model are iteratively attached to the gene selected in the previous step at N=1. The estimation of all procedures listed before in the case of single-feature models is compatible with the workflow, including multi-feature models. There is one additional stop criterion at this level. This criterion is also reached if all features are included in the created model at the highest level of complexity. When all possible models for specified complexity levels are calculated, the model with the lowest BIC value is selected, following the procedure described earlier for single-factor models. At this stage, when the level of complexity satisfies the assumption that N>1, an additional part of the tools' work appears, enabling the comparison of models of varying complexity. The introduction of the Bayes Factor (10) metric makes it possible to compare models with a complexity of N-1 and N.

$$BF = e^{(LL_{N-1} - LL_N)} \quad (10)$$

If the value of the BF metric is greater than $10^{1.5}$, the model with a higher degree of complexity is more profitable. Therefore, the implemented algorithm builds models with successive degrees of complexity. The model-building procedure is completed when one of the two algorithm termination criteria (smaller

BF value at the next level of complexity or the highest possible complexity level) is reached. Below, in Figure 2, a simplified diagram of the implemented machine learning algorithm is shown based on logistic regression methods.

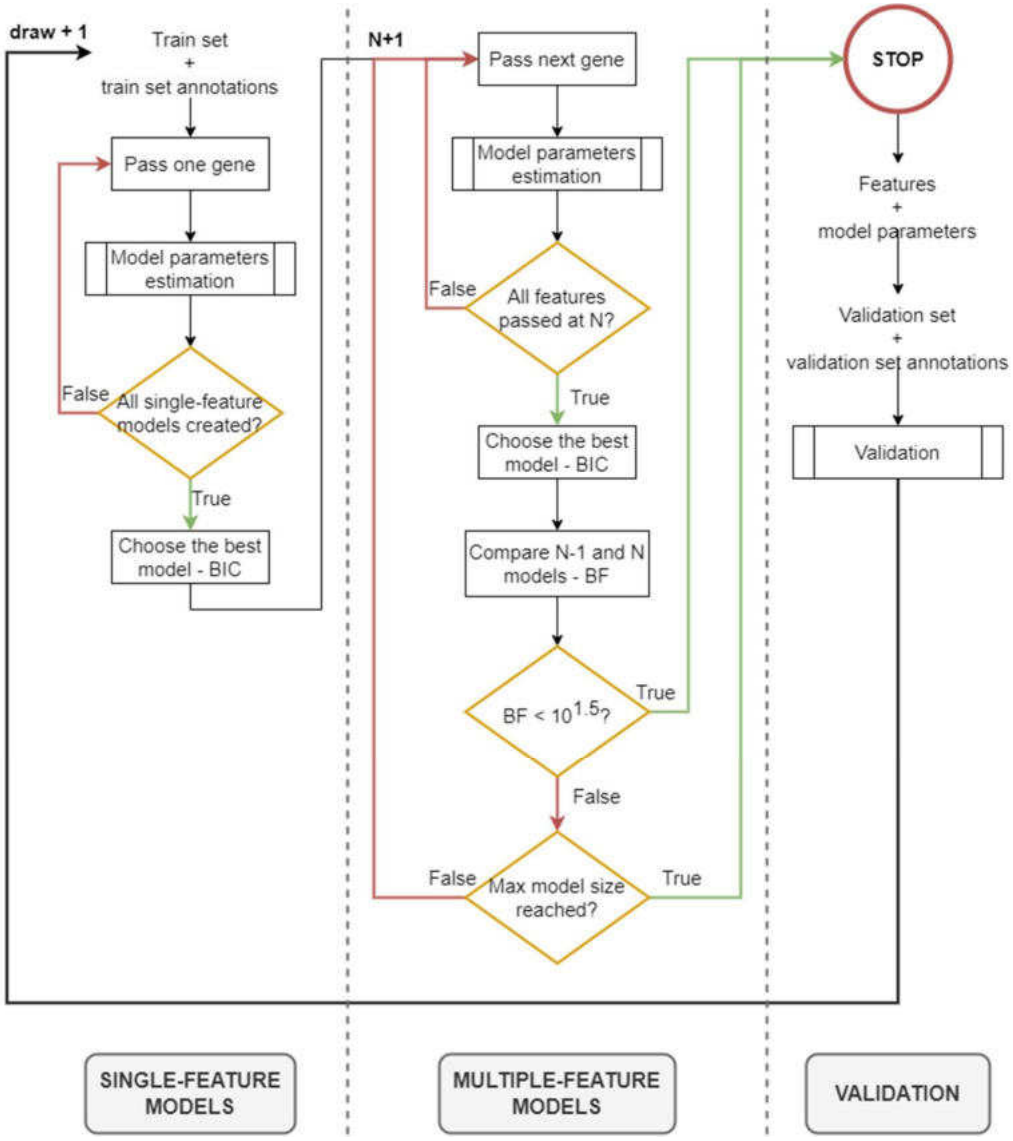


Figure 2. Outline of the implemented algorithm for the features selection and cells' classification purposes.

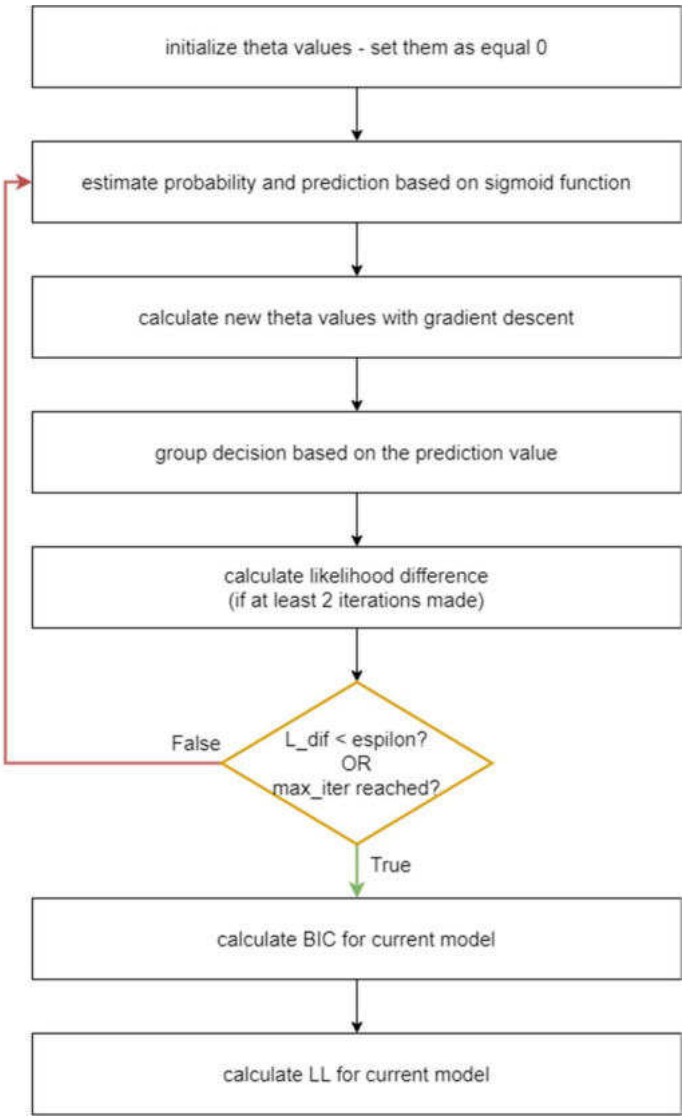


Figure 3. Outline of the implemented algorithm for the features selection and cells' classification purposes – model parameters estimation process details.

After building the entire workflow architecture, it became necessary to optimize the full implementation. Optimization was required due to the extended running time of the algorithm. This procedure was carried out based on the selected set of cells and genes from the set B data. The optimization was based on the first indexed 300 cells and 200 genes. The following improvements were made with the original version of the program: hiding redundant information displayed in the console, vectorizing the code (only possible to vectorize elements), and changing all analyzed data frame objects on the matrix structures. The actions taken made it possible to shorten the number of rows for the implementation of the algorithm by approximately 33% of the original implementation version. More importantly, the implementation improvements also reduced the algorithms' uptime significantly. Table 3 compares the algorithms' working time for one and five built models.

Table 3. Comparison of the runtime of the raw and optimized versions of the implemented algorithm.

Implementation	1 Model	5 Models
Raw algorithm [sec]	90.99	340.59
Optimized algorithm [sec]	18.78	52.13
Gain of time [%]	79.36	84.69

The *tic()…toc()* function implementation in the *tictoc* library from the R environment was used to determine the algorithms' operation time. This basic optimization of the implemented algorithm allowed for a significant improvement in the program operation times.

4.2.2 Neural networks-based workflow

The second approach to modeling the gene signature of irradiated cells, used in the dissertation, is based on machine learning algorithms using neural networks. To create the network structure, the *Sequential()* function from the *tensorflow.keras* library [56] was used. This function makes it possible to design and develop the scheme of a neural network model very transparently, including the number of network layers used, the number of neurons in individual layers, and the neuron activation functions used in specified layers.

In the case of this dissertation, the sigmoid activation function was considered. The task of the activation function, as the name suggests, is to decide whether a given neuron in the neural network should be activated. Activation means determining whether the information provided by this network element is relevant to its prediction process. The task of this function is, therefore, to transform the summary information, in the form of the transmitted signal value x_i, weights for the input neuron w_i, and bias b, flowing into a specific neuron into an output value transferred from this neuron to the next layer of the neural network. However, using a considerable simplification about the activation function's purpose introduces non-linearity in the analyzed and utilized network. This nonlinearity is necessary to build the least complicated network to learn the complex scheme of the analyzed data. In neural networks, it is possible to use many different types of activation functions, but the most commonly used is the sigmoidal function [57] which is defined by (11):

$$f(z) = \frac{1}{1 + e^{-z}} \tag{11}$$

Where:

$$z = \sum_{i=1}^{N_{input}} x_i \times w_i + b \tag{12}$$

This function transforms the appropriate input values into the range from 0 to 1. It is widespread in currently used solutions using neural networks due to the content of resulting values that can be directly translated into probability values. A significant part of the problems for which machine learning methods are used is based on probability. Another reason for the frequent use of the sigmoidal activation function is its smooth transition between the result values, which limits the occurrence of jumps to the following result values [57].

Essential elements of each neural network are built-in layers containing a certain number of neurons. There are generally three layers: input, hidden, and output [58] [59]. The neurons building the input layer transmit raw information about the introduced features to the next layer, the hidden layer. What is more, in a complete neural network model, there may be several hidden layers depending on the needs or specificity of the data. It is in the hidden layers that the conversion of individual signals takes place, resulting in the input information to the last output layer. The received signal is transformed into the expected results in the output layer. Figure 4 shows a simplified diagram of a fully connected neural network.

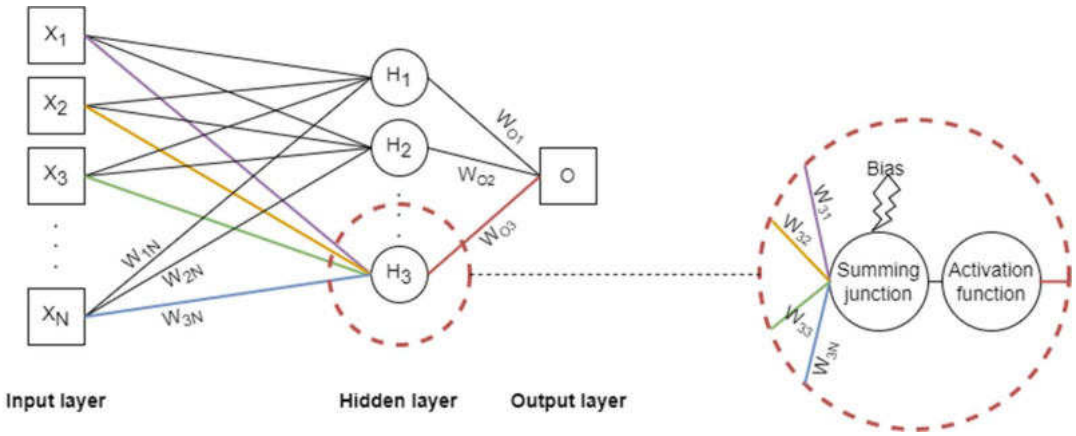


Figure 4. A simplified diagram of the neural networks structure.

The next worth concerning concept in modeling, particularly in the case of neural networks, is the flow of information. The forward and backward propagation procedures play a vital role in this phenomenon. Forward propagation is data flow from the input layer to the output layer. The defined hidden layers calculate the input information using feature values and specified weights in this procedure. The resulting values from all included hidden layers are then transformed in the output layer to obtain the result. In the case of this procedure, the activation function is responsible for converting the signal coming out of the neuron, which in combination with signals from other neurons, becomes the input information to neurons located in the next layer of the network. On the other hand, backward propagation converts the obtained results from the output layer through successive hidden layers toward the input layer. The weights of individual neurons are recalculated in such a way as to minimize the difference between the output vector from the neural network $\hat{y}$ and the expected (true) output vector y. This procedure reduces the cost function by adjusting the weights and biases that build the neural network. Therefore, the loss function plays a significant role in the flow of information. It is generally possible to update the model's parameters by utilizing the loss function. It informs about the essential differences between the predictions and the correct result, which we want to get as close as possible. Most commonly used for two class problem is the binary cross-entropy (BCE) function, according to equation (13):

$$BCE(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \times \ln(\hat{y}_i) + (1 - y_i) \times \ln(1 - \hat{y}_i) \quad (13)$$

Following the procedures described, the conversion of model parameters was performed using the *fit()* function from the *tensorflow.keras* library in the Python environment. Moreover, the Adaptive Moment Estimation Adam [59] optimizer was used. It is nowadays one of the most frequently utilized optimization methods for neural network purposes [60]. Adam is often used to find the individual learning rate values for each parameter and to update weights and biases available in the neural networks-based model. This method of optimizing the parameters of neural network models is popular primarily because it is computationally efficient and very well suited to working with large datasets in the context of the number of features [61].

5 Preliminary data analysis

The preliminary single-cell sequencing data analysis workflow is fundamental to pre-investigate the structure of the real data. This part consists of a series of procedures subject to the quality control of cells and genes and visualization procedures. The last mentioned are extremely important in studying the internal structure of data, its volatility, and dependencies. These methods very often detect structures hidden inside the analyzed data. Thus, it is essential to carefully examine the cells and genes' quality to be subjected to the complicated procedures of discovering a genetic profile of irradiated white blood cells. The preliminary analysis procedure was applied to samples from the ex vivo experimental environment, both the control and the irradiated, with a dose of 1 Gy samples.

5.1 Data pre-processing

The data from two technical repetitions of the experiment carried out in an ex vivo environment, as mentioned in the *Material* section, included five group designations of cells: control, irradiated, *mixed*, *multiplet*, and *undetermined*. Three groups of cells without explicit control or irradiated group assignment (*mixed*, *multiplet*, and *undetermined*) were filtered out before proceeding with the proper quality control procedures. Due to the described filtration rule, 1516 cells were removed from further analysis in the case of set A, and 2633 cells were rejected from the set B data matrices.

Before the proper quality control analysis, the second step was to check the counts across the control and irradiated samples for the well-known GAPDH housekeeping gene (HKG). It was found that although the GAPDH gene is widely regarded as an indicator of quality or a gene of high utility value in data normalization, it cannot be used as a determinant for data derived from single-cell sequencing experiments. Table 4 gives detailed information on the number of zero-count cells for this gene over both technical repetitions of the ex vivo experiment.

Table 4. Number of zero-count cells for GAPDH gene for ex vivo datasets.

	Set A	**Set B**
Control	381 (24.05%)	795 (34.55%)
Irradiated	308 (27.04%)	675 (33.95%)

Removal of a described number of cells detected as invalid (about 1/4 of set A and about 1/3 of set B data) means losing a lot of potentially valuable information carried across the datasets. Due to the in-depth analysis of literature sources, it turned out that studies on the stability of genes, once considered an excellent indicator, clearly show that the GAPDH gene should not be regarded as HKG for sample normalization purposes due to the bimodality of the distribution [62] [63] [64]. For this reason, further analyses related to the GAPDH gene were abandoned in this dissertation, and attention was drawn to another gene that can be used as the HKG. The HLA-A gene was next recognized as potentially valuable for the described problem and analyzed for zero-count cells. For set B, only six cells from the irradiated sample were found to have zero counts for this gene. These cells were filtered from the dataset for further quality control purposes. This procedure allowed the filtering of cells for which proper functioning was not maintained during the experiment. In addition, it was found that the selection of an appropriate HKG is not always evident in the case of single-cell data. The detected discrepancies for the GAPDH gene as a housekeeping gene, in the case of data from single-cell sequencing experiments, have been fully confirmed by numerous literature sources [62] [63] [64].

The proper quality control consists of three steps covering issues: gene quality control based on Unique Molecular Identifiers (UMIs) distribution and cell quality control based on library size and expressed features. Additionally, in two step-manner, the number of genes with zero counts over all observations was assessed to filter out cells without significant analytical value.

For ex vivo data analysis, the UMI information about the genes was attached in the additional data quality file created at the data collection and preparation stage. Unique Molecular Identifiers are random oligonucleotide barcodes used in high-throughput sequencing techniques. Including UMI in the exact location in every fragment of genetic material during library preparation but before PCR amplification makes it possible to distinguish between PCR duplicates that have identical UMI sequences [65]. As a result, each of the genes in the panel was described by one of the three gene statuses: *not detected*, *low depth*, and *pass*. As mentioned before, the *not detected* status means that the gene was not detected due to zero reads despite its presence in the gene panel. The *low depth* status means the minimum sequencing depth has not been reached. Both described statuses indicate significant abnormalities in determining the counts of individual genes. *Not detected* and *low depth* described genes were filtered from both ex vivo data sets. Subsequently, 42 and 28 genes were removed from further analysis in sets A and B. At this step, genes with zero counts across available cells were also checked in the individual ex vivo dataset. For this reason, four genes from set A were rejected, and two from set B. Summing up this step of the quality control workflow, 30 genes were filtered from set A, and 46 poor-quality genes from Set B.

The library size is another essential property used concerning the analyzed data sets, and it is the total sum of the counts across all available features. Filtering based on library size allows identifying cells on which RNA capture in the library preparation process (cDNA conversion and amplification) was inefficient. Therefore, these cells are characterized by small library sizes. For the analyzed single-cell sequencing data, the cumulative sum of the counts across all available genes was calculated for each cell. Thus, the library sizes of each cell were determined. From the data prepared this way, histograms of the library sizes of individual cells were drawn for both technical replicates of the ex vivo experiment. The median absolute deviation (MAD) criterion derived from (14) was used to determine the appropriate cut-off point for acceptable-quality cells.

$$MAD_{threshold} = M(LS) - 3 \times (M|LS_i - M(LS)|) \quad (14)$$

Where:
M is the median value,
and *LS* is the library size.

As described, MAD measures the variability of a univariate sample of quantitative data. Utilizing this measure, the deviations of a small number of outliers are insignificant. Library size histograms for set A and set B data, along with the cut-off threshold values for acceptable and poor-quality cells, are presented in Figure 5 and Figure 6, respectively. The given measures have been scaled to $\log_{10}$(library size) values for better visualization.

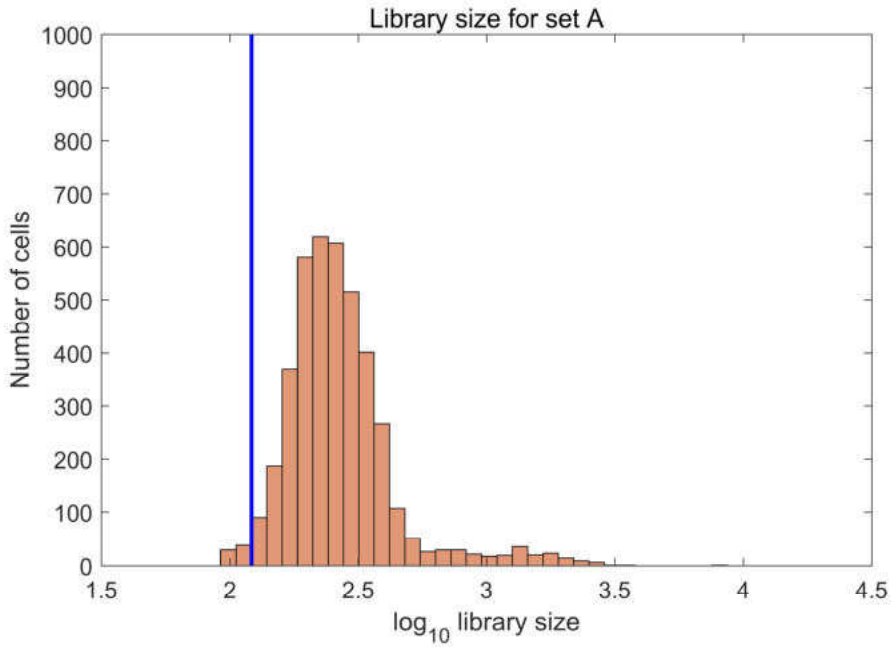


Figure 5. Histogram of library sizes with marked $\log_{10}$MAD=2.08 threshold for set A.

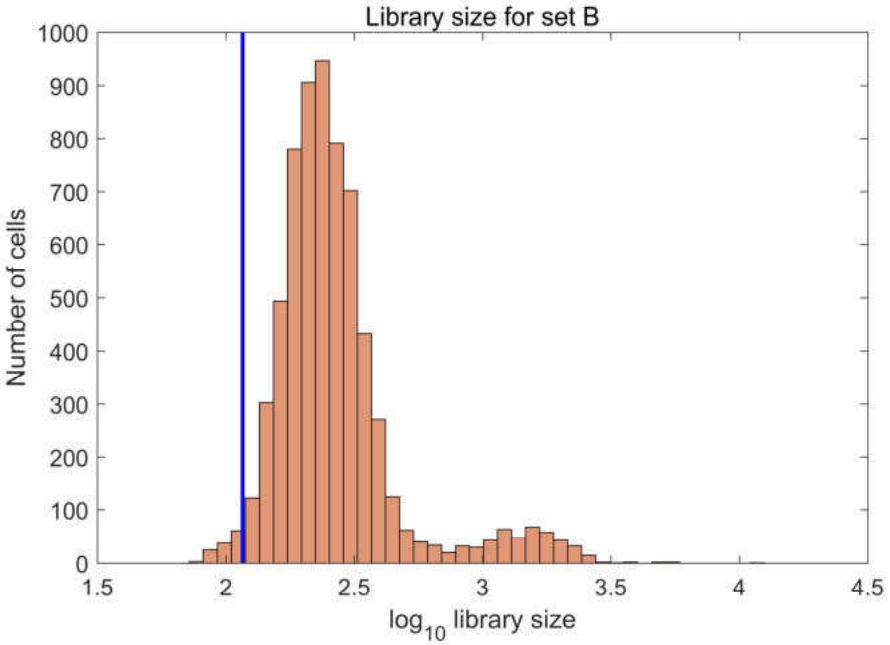


Figure 6. Histogram of library sizes with marked $\log_{10}$MAD=2.07 threshold for set B.

The minimum and maximum library size values for set A were equal to 92 and 8654, respectively; for set B, these values were equal to 72 and 12448, respectively. Cells with an estimated library size value less than the marked $MAD_{threshold}$ were rejected from further analysis using the designated thresholds. The library size values of individual cells marked as low-quality cells are to the left of the specified

threshold values in Figure 5 and Figure 6. Cells marked as low-quality cells were filtered from the data sets: 67 low-quality cells were filtered from set A, and 114 low-quality cells were filtered from set B.

The next analyzed measure was the number of expressed features in each cell. It is defined as the number of genes with non-zero counts for specified observation. Cells with very few genes in expression (non-zero counts) are treated as poor-quality cells. For both technical repetitions of the ex vivo experiment, the number of features in expression was calculated for each cell as the sum of the features with non-zero counts. This metric's minimum and maximum values were for set A, equal to 29 and 144 features in expression, and set B, equal to 31 and 169 features in expression. Tukey's criterion for extreme outlier values was used to determine the cut-off point for this metric. Equations (15) present the formulas for the lower and upper thresholds for the number of expressed features.

$$Tukey_{lower} = Q_1 - 3 \times IQR$$
$$Tukey_{upper} = Q_3 + 3 \times IQR \quad (15)$$

Where:
Q_1, Q_3 are 1st and 3rd quartiles,
and IQR is the interquartile range.

Figure 7 and Figure 8 show the histograms of the number of expressed features with marked $Tukey_{threshold}$ values for the set A and set B data, respectively. Moreover, the histograms are presented on a $\log_{10}$(expressed features) scale for better visualization. In both histograms, the median values for the number of features in the expression for both technical replicates were also marked with a solid red line. The median values were 1.80 and 1.81 for sets A and B, respectively. In the case of the set A data, two cells were found outside the designated threshold values (one cell was above, and one cell was below the selected lines). These two cells were filtered out from the mentioned set A data.

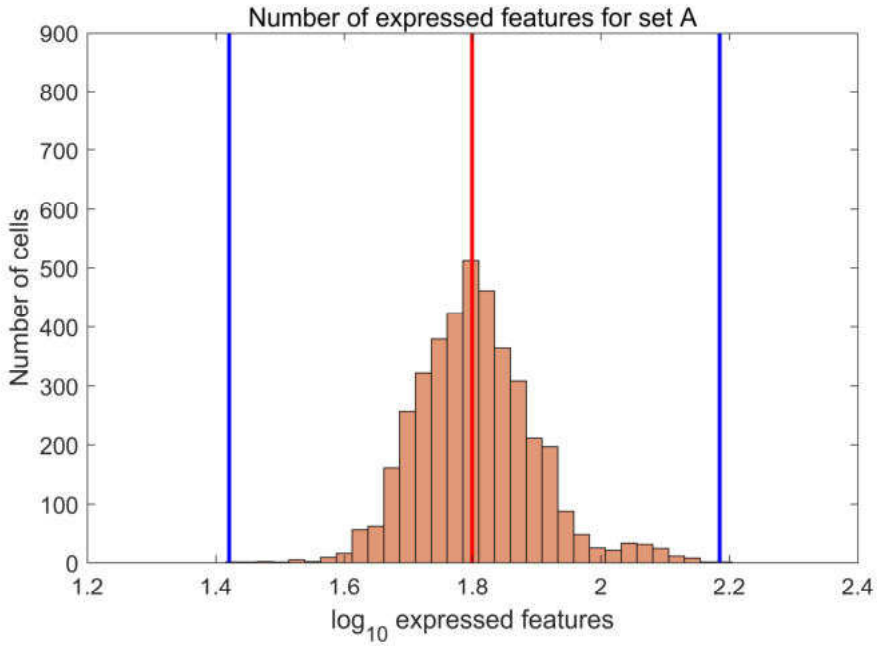


Figure 7. Histogram of the number of expressed features with marked $\log_{10}Tukey_{lower}$=1.42 and $\log_{10}Tukey_{upper}$=2.18 thresholds for set A.

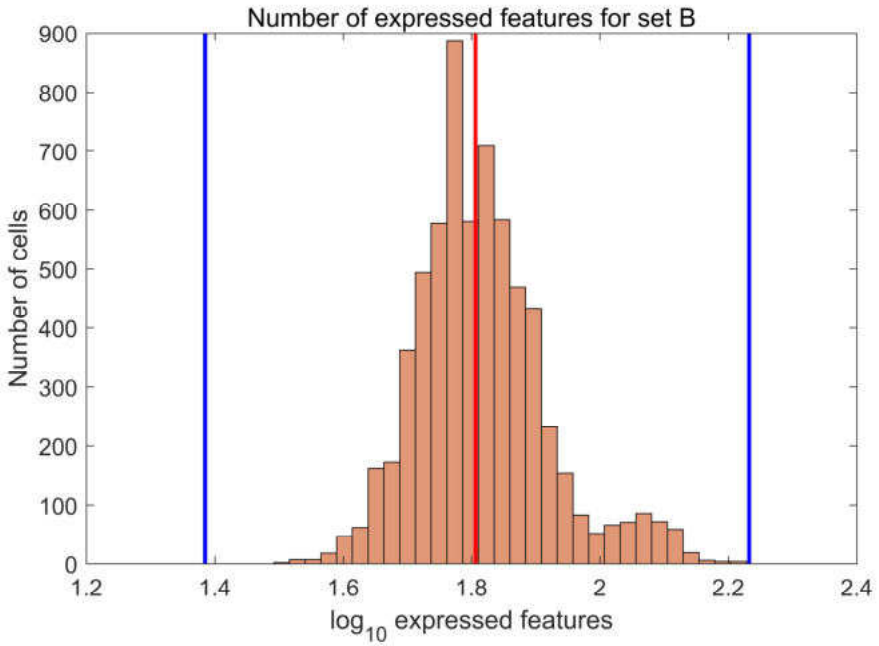


Figure 8. Histogram of the number of expressed features with marked $\log_{10}\text{Tukey}_{\text{lower}}$=1.39 and $\log_{10}\text{Tukey}_{\text{upper}}$=2.23 thresholds for set B.

For the number of expressed features, a second metric was also used for threshold values estimation concerning good-quality cell recognition. As with the described library size metric, the MAD measure was used to determine the threshold values. In the case of both ex vivo datasets, the determined $\text{MAD}_{\text{threshold}}$ values indicated the need to remove 66 cells from set A and as many as 106 cells from set B. Due to the possible significant loss of acceptable-quality cells, an additional and detailed analysis was performed. The number of expressed features distribution division was utilized in the Gaussian mixture models (GMM). To estimate the best number of components, there was used the Bayesian Information Criterion. The analyzes carried out this way for both datasets, with utilized BIC criterion, indicated the selection of three GMM components. Histograms of the number of expressed features with a marked $\text{MAD}_{\text{threshold}}$ and designated GMM components are presented in Figure 9 and Figure 10. For both datasets, the cells to the left of the selected threshold values constitute a significant proportion of one of the GMM components. Therefore, they cannot be considered poor-quality cells. Due to this, no cells from both ex vivo datasets were removed from further analysis in this step.

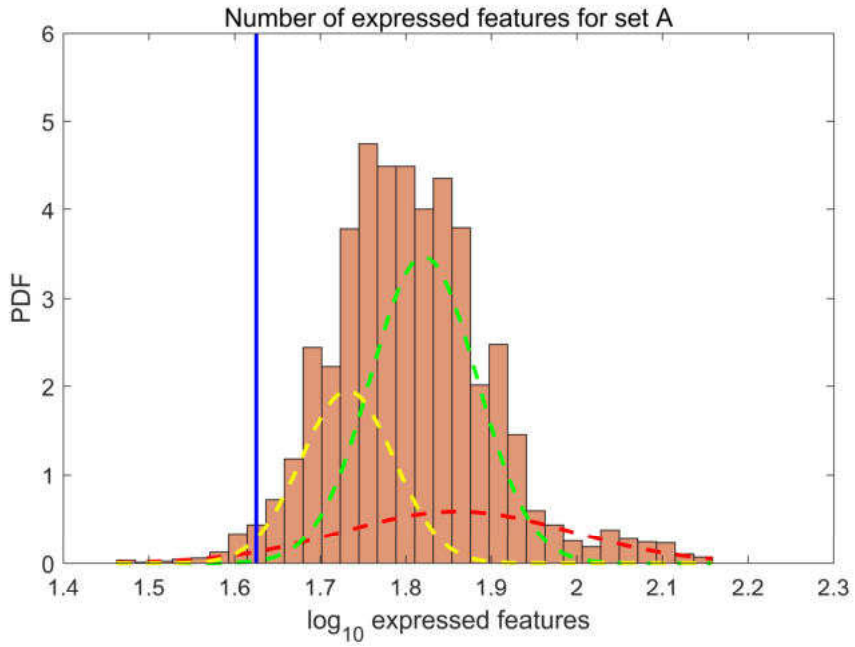


Figure 9. Histogram of the number of expressed features with marked $\log_{10}$MAD=1.63 threshold and GMM components (red, yellow, green) for set A data.

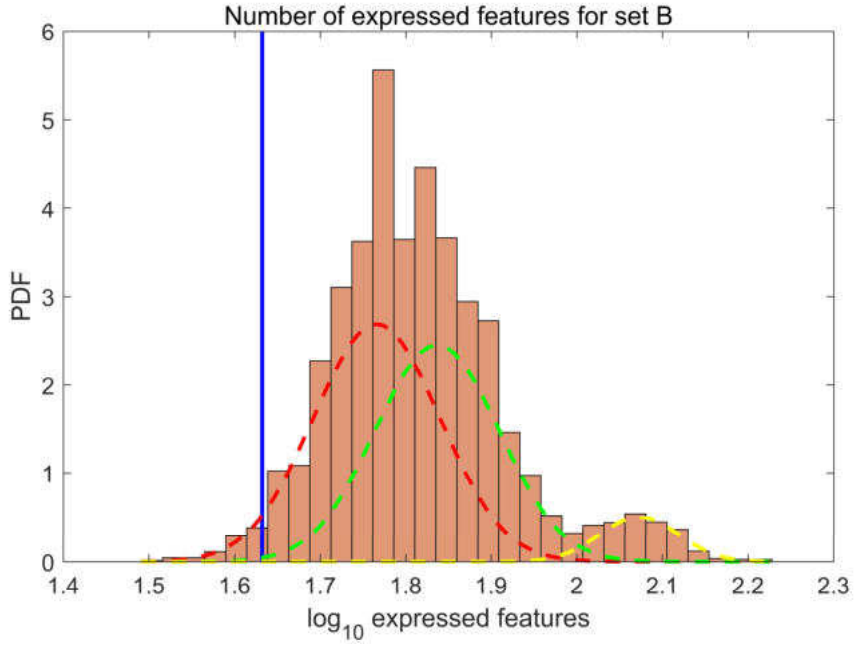


Figure 10. Histogram of the number of expressed features with marked $\log_{10}$MAD=1.63 threshold and GMM components (red, yellow, green) for set B data.

As the last part of the QC workflow, based on the data from ex vivo experiments, the informativeness of the genes remaining in the analysis was checked. For this purpose, the number of cells with non-zero counts for specified genes was found. Features with non-zero counts in less than three cells were filtered from the analysis. Based on this criterion, ten genes were removed from set A, and 16 genes were extracted from set B. A summary of the quality control performed on genes and cells for both technical repetition datasets is presented in Table 5.

Table 5. Summary of the quality control workflow performed for set A and B data, specifying the number of genes and cells.

		Set A	Set B
Removed	Control cells	25	49
	Irradiated cells	15	26
	Cells (total)	40 (1.47%)	75 (1.75%)
	Genes	56 (12.39%)	46 (10.18%)
Remaining	Control cells	1559	2252
	Irradiated cells	1124	1962
	Cells (total)	2683	4214
	Genes	396	406

As a result of the preliminary data analysis and quality control, there were removed poor-quality cells and genes from the analyzed ex vivo datasets. There are still many valuable features left in the analysis: for set A, there are 396 genes, and for set B, there are 406 genes of satisfactory quality and informativeness. For the quality control performed on cells and summing up the total number of cells still available in the analysis, there are 2683 cells left for set A and 4214 cells for set B data.

5.2 Dimensionality reduction and visualization

Visualizing such complex, multi-dimensional single-cell sequencing data requires prior dimensionality reduction. The visualization process must be carried out within an acceptable time window while maintaining the best reflection of the raw data structure. Analyzing hundreds of features simultaneously is a computationally highly complex process that takes much time. The purpose of visualization is, first of all, to get acquainted with the data structure and to look for hidden connections of features, if any are present. Of course, when reducing dimensions, special attention should be paid to maintaining important information transferred within the available features. The dimensionality reduction for data visualization purposes cannot be a priority stage; it should be an auxiliary one, allowing for the reduction of computational costs. The $\log_2(\text{counts}+1)$ transformation was performed on the analyzed data sets to represent the features better. Adding the unit value to the number of counts will not worsen the representation of the analyzed data but will avoid zero-count transformations. It will primarily allow measuring changes in expression (counts), which is a much more interesting biological phenomenon.

First, the dimensionality of the data sets was reduced using Principal Component Analysis (PCA) procedures. Importantly, with the use of PCA, it is not possible to return to the original structure of features. PCA performs a series of transformations, changing individual features into their combinations, thus creating entirely new features called PCA components. This is an exciting approach regarding a significant reduction in dimensionality while maintaining vast information from the raw features. However, if further research requires a return to the space of primary features, other dimensionality reduction methods should be used, supporting the feature selection problem. In the described case, dimensionality reduction was used only for visualization purposes. Therefore, there is no need to return to the primary space later, and a fundamental goal is to reduce the number of features

with the lowest possible loss of information carried by them. In such a complex system, the linkages of individual genes can have significant signatures. Based on PCA procedures, the methodology used is a very conservative choice, leading to a substantial reduction in the dimensionality of data sets. It assumes the selection of only those PCA components for which significant changes in the explained variance of the set can be subsequently observed. Figure 11 and Figure 12 show the percentage of explained variance of the datasets for individual components, with the number of PCA components selected for further visualization procedures marked successively for sets A and B. The graphs show only the first 100 PCA components for readability purposes.

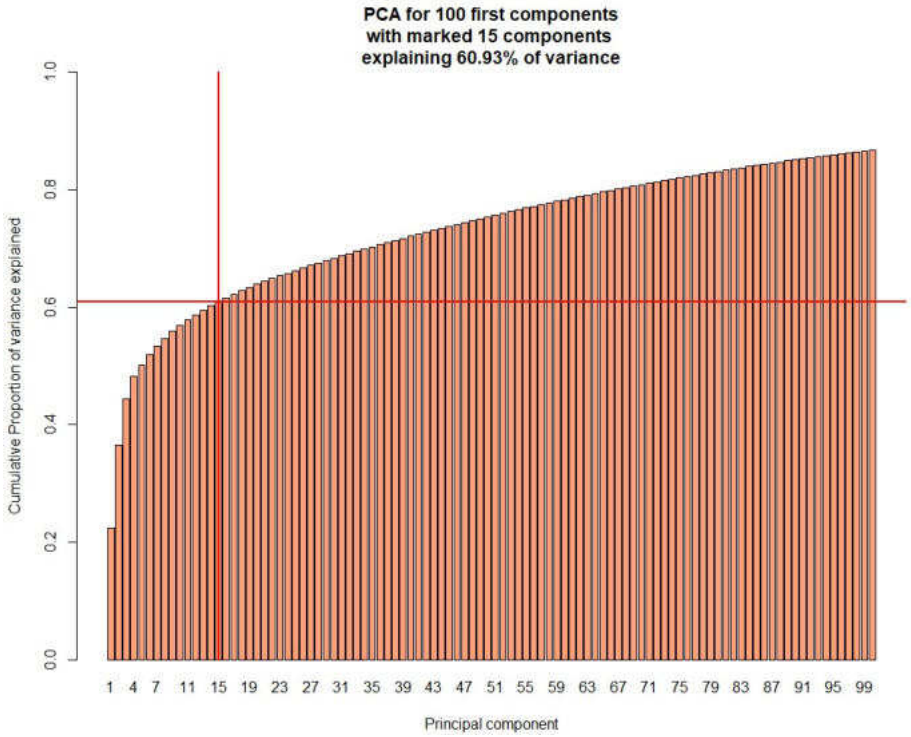


Figure 11. PCA components with the marked threshold for a number of chosen PCA components for set A.

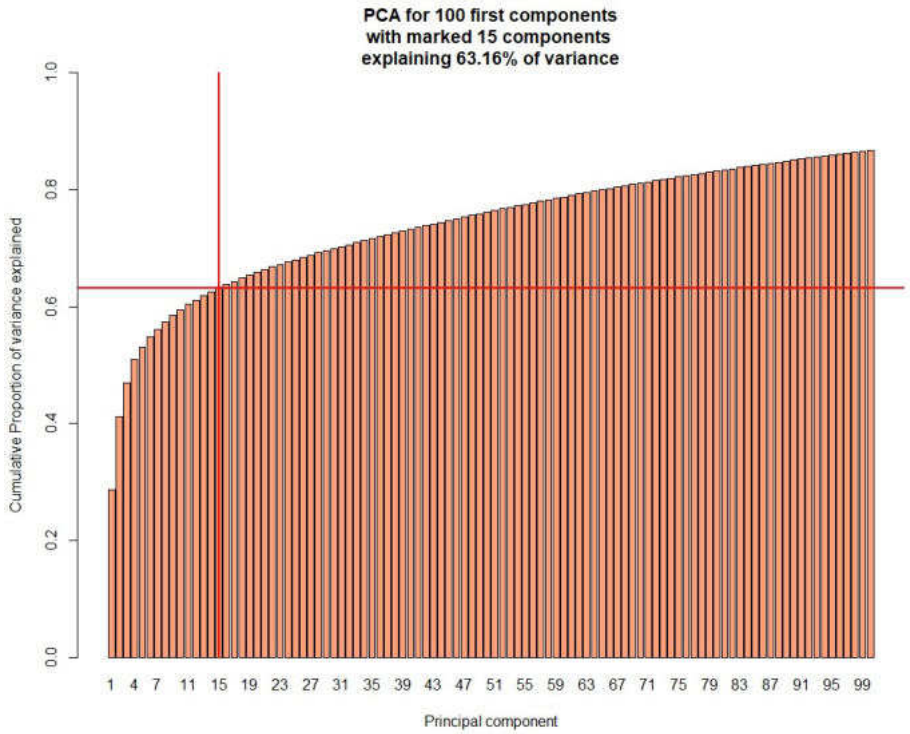


Figure 12. PCA components with the marked threshold for a number of chosen PCA components for set B.

To visualize spatially reduced feature sets, the first 15 PCA components (out of 396 available features) were selected for set A, explaining almost 61% of the variance. In the case of set B, the first 15 PCA components (out of 406 available features) were selected, explaining over 63% of the variance.

The UMAP tool was used based on the chosen PCA components on dimensionality-reduced datasets. The visualization was performed using the unsupervised approach to enable the detection of possible structures within the data, independent of the assignment of individual cells to the control or irradiated sample. An appropriate color legend was also assigned to the spatially arranged cells to define their original affiliation. Results for both datasets are presented in Figure 13 and Figure 14.

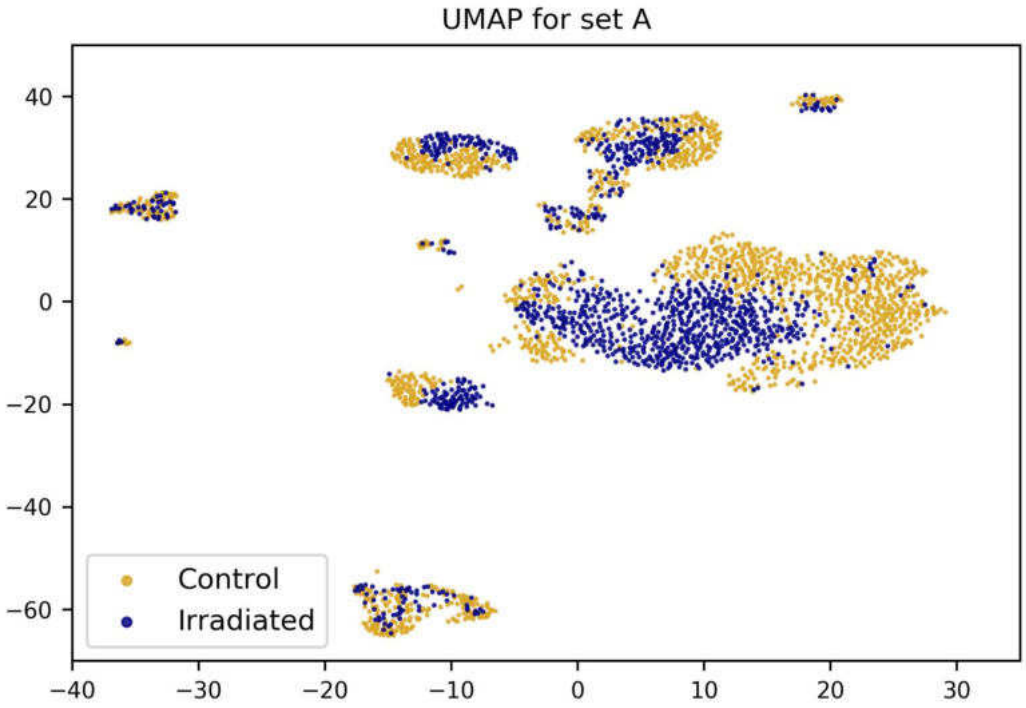


Figure 13. UMAP unsupervised representation for set A.

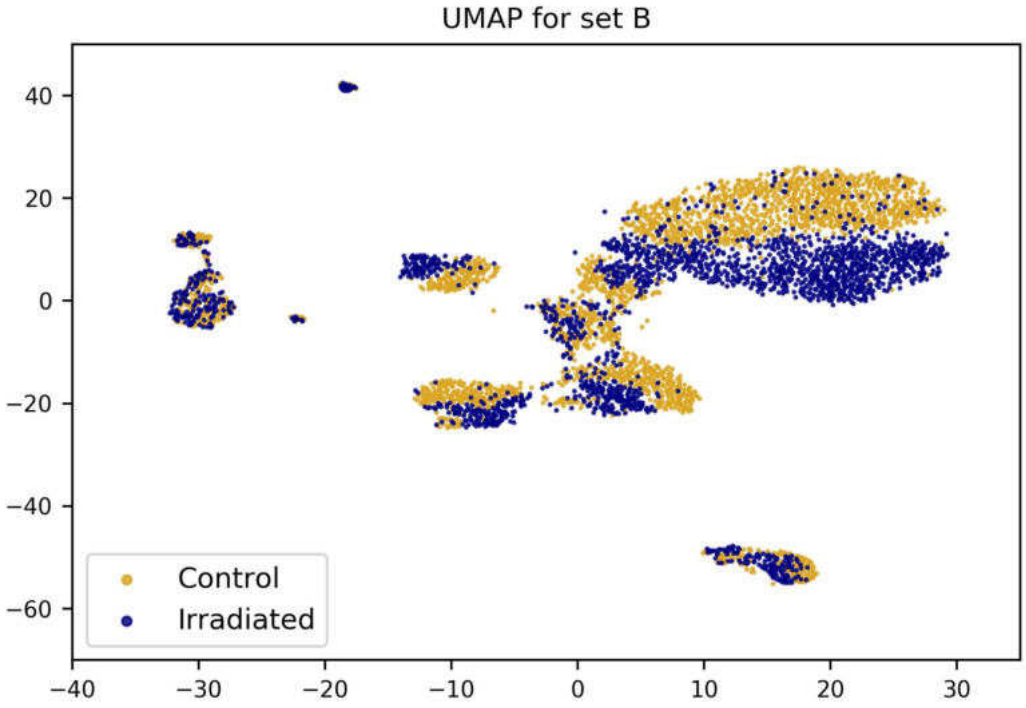
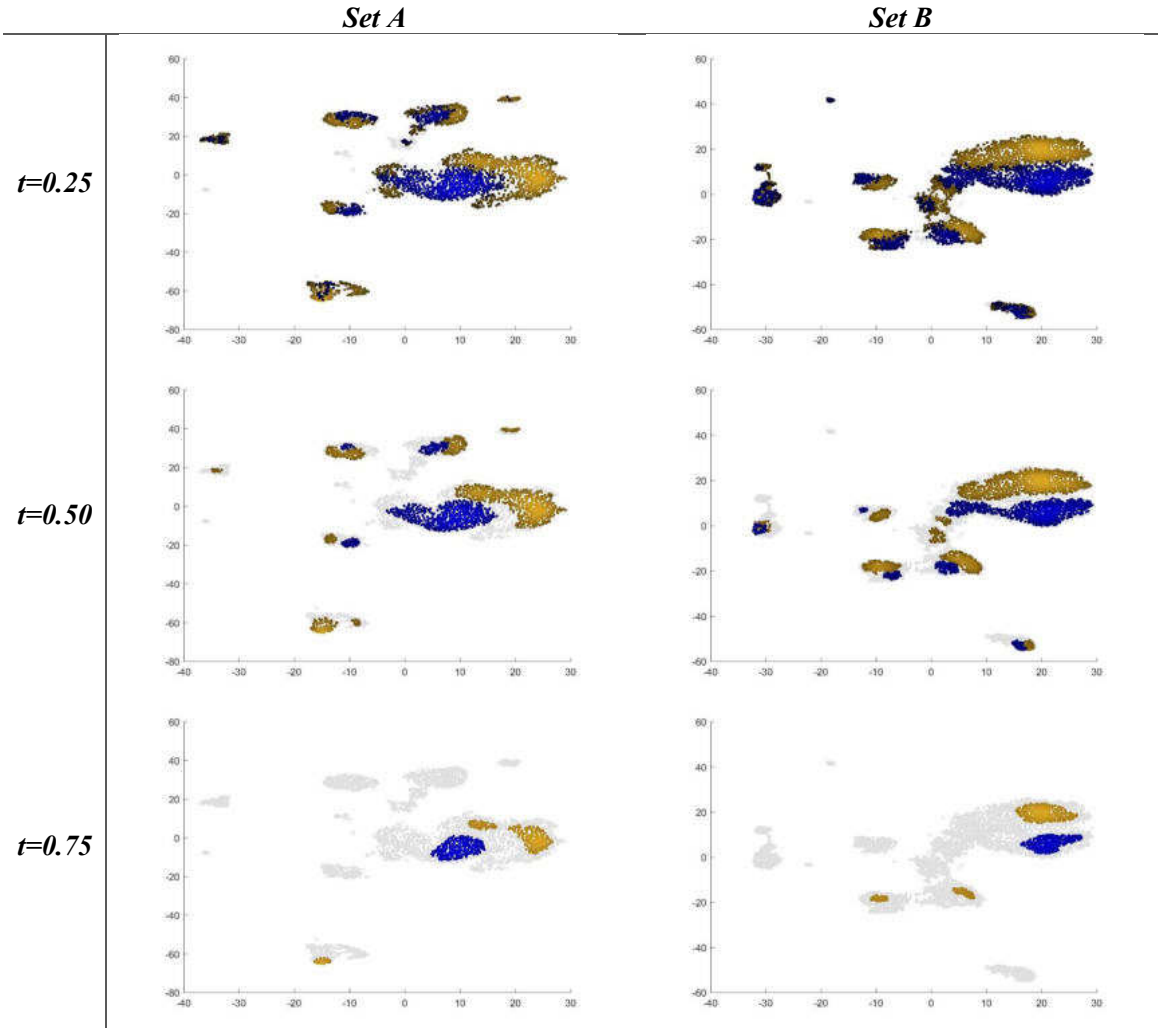


Figure 14. UMAP unsupervised representation for set B.

Based on the two-dimensional UMAP projections, separated clusters of observations for analyzed datasets can be easily delineated. Each collection consists of both control and irradiated cells. Thus, the separation of individual clusters is not caused by the presence of two cell samples but by an effect much more vital than the radiation factor. It is worth emphasizing that these clusters were detected using an unsupervised approach. It means visible clusters were separated without providing information about the cell memberships to the control or irradiated sample. Therefore, this analysis made it possible to conclude that the hidden data structure is related to the high heterogeneity of the studied datasets. It undoubtedly dominates over the radiation factor in the unsupervised approach. To better identify the distribution of control and irradiated cells in the separated clusters, the maps of the highest concentration of these cells in the 2-dimensional UMAP space were analyzed. Table 6 shows the results of the cell-cluster analysis for the three different thresholds for testing cell density arrangement.

Table 6. Results of the analysis of the cells' density inside the designated clusters for different concentration thresholds *t*.



As a result of the analysis of the clusters' density arrangement, it can be clearly stated that the centers of clusters for control and irradiated cells, despite occupying the same separated clusters, have different locations. The conducted analysis allows for drawing two basic conclusions. The first undoubted finding is the hidden structure of the analyzed ex vivo datasets related to their high heterogeneity level. Leaving this problem unresolved in building genetic profiles of irradiated cells and classifying these cells may negatively impact further considerations. It is, therefore, necessary to discover the cause of the heterogeneity of these datasets to make the required manipulations to monitor this effect. The second conclusion is the apparent influence of radiation on the differentiation of control and irradiated cells. The analysis of cluster density made it possible to detect that despite the superficially visible blurring of differences between these groups of cells inside the designated clusters, these cells group together in other locations of the presented 2-dimensional UMAP space.

5.3 The hold-out test structure

Preparation of the test data structures for both technical repetitions of the ex vivo experiment is enormously essential to maintain the possibility of testing solutions on the data set that did not participate in the individual stages of building the gene signature or, above all, models' learning procedure. A part of the A and B datasets was set aside to ensure the complete independence of a particular pool of observations from the processes of creating the genetic profile of irradiated cells. These sets will be consistently called the test sets in the doctoral dissertation. They were created by randomly selecting 20% of the data remaining after the quality control process. Notably, the ratio of control to irradiated cells in the test sets was retained to reflect the composition of the initial input data. The test sets remain the same, considering their design, throughout recognizing the genetic profile of irradiated cells, regardless of the machine learning methodology used. Table 7 shows the quantitative compositions of individual cell samples for the randomly selected test and model structures.

Table 7. The composition of the test and model structures.

		Set A	**Set B**
Test structure	Control	306	452
	Irradiated	231	391
	Total	537	843
Model structure	Control	1253	1800
	Irradiated	893	1571
	Total	2146	3371

The distribution of the test set observations is shown in Figure 15. For visualization, there was utilized the UMAP tool. It was based on previously described 15 PCA components data for all observations available after the quality control in set B.

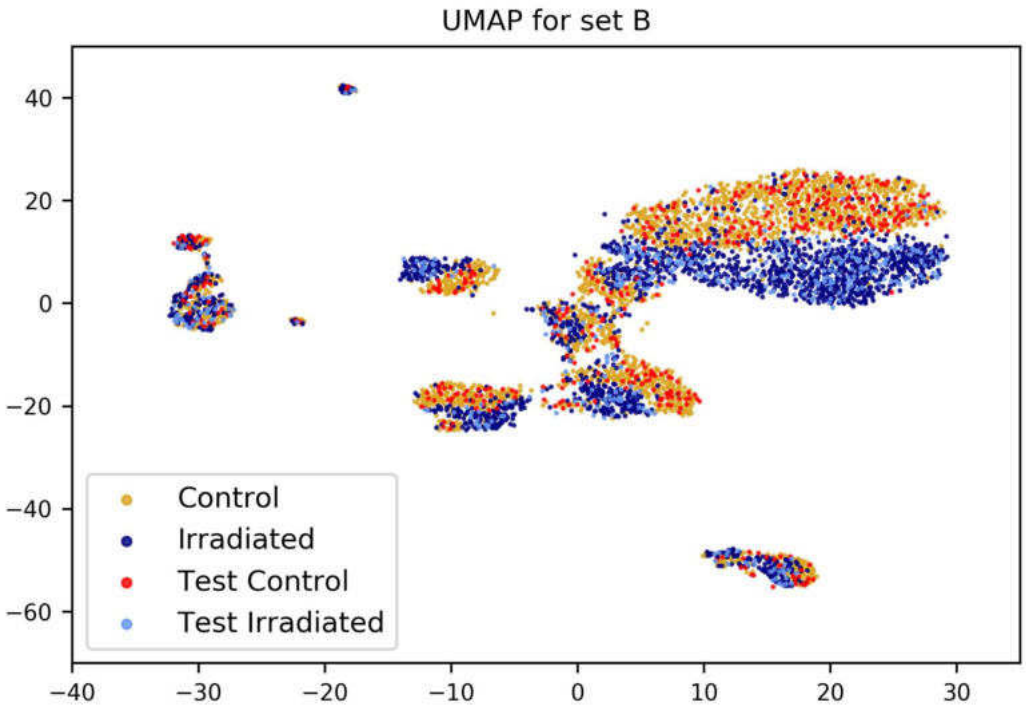


Figure 15. Distribution of the test set over the B dataset.

According to the assumptions, the random selection of observations for the test set made it possible to ensure an even representation of cells in the 2-dimensional space of the B dataset. The presented points representing the control (red) and irradiated (light blue) cells of the test set cover the area of the drawn observations of the model structure for the control (goldenrod) and irradiated cells (navy blue). Moreover, there are no places of particular accumulation of cells of the test set in specific parts of the presented 2-dimensional UMAP space. This guarantees a full and independent representation of the cells for the test set and will allow the testing and direct comparison of classification results for different approaches used.

6 Ex vivo irradiated cells' genetic profile recognition based on the logistic regression methods

The ex vivo data analysis, in the form of two technical repetitions of the experiment, sets A and B, concerns the set of features and observations after data quality control. This part of the dissertation is intended to determine the ex vivo irradiated cells' gene signature. Based on selected genes, it will be possible to distinguish between control and irradiated cells in the mixed-cells-membership datasets. The purpose is to find features indicating whether the analyzed cells among the white blood cell sample were irradiated. The following subsections exhaustively describe the research workflow, leading to determining the gene structure of irradiated cells based on logistic regression methods.

6.1 Irradiated cells' genetic profile recognition based on the white blood cell dataset

The irradiated cells' genetic profile identification stage using logistic regression methods is based on data set B. In the next step, there was an independent testing of the model based on data set A, derived from a technical repetition of the same experiment. The developed workflow consists of the features selection using modeling based on logistic regression methods, determining the significance of features based on the generated lists of features using the appropriate metric, final model building, and mentioned independent testing of the model.

The feature selection using LR methods, implemented according to the proposed workflow, is described in detail in the section *Logistic regression-based workflow*. The algorithm was designed to build 50 complete models based on the provided set B model data structure. There was an obvious need to generate train and validation sets for the LR-based model learning. When the algorithm is run once, it is possible to generate only one set of selected features; therefore, to obtain 50 assumed models, the algorithm was run 50 times. The sets generation system was programmed at the entrance to the models' implementation. The graphical presentation of the randomization of individual datasets is presented in a simplified way in Figure 16.

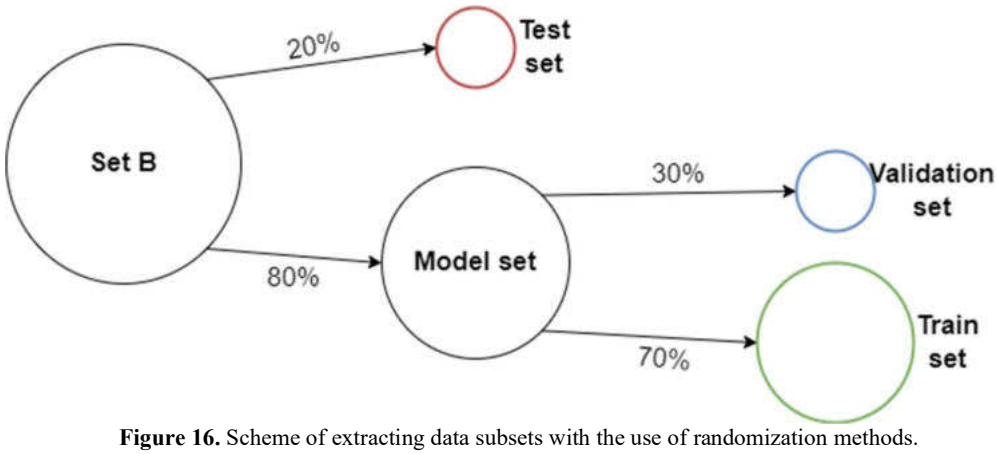


Figure 16. Scheme of extracting data subsets with the use of randomization methods.

The task of the sets generation system is to draw a validation data structure consisting of 30% of the input cells available in the model data structure (remaining after the test set was separated). Moreover, the validation set was always composed of the same number of control and irradiated cells to ensure the balanced composition of this set. The remaining part of the model set formed the training structure. Each launch of the model was associated with the analysis of the train and validation data structures, modified in terms of cells' composition.

As a result of the features selection process utilizing a self-learning LR-based algorithm, 50 sets of features included in the created models were obtained. For each model, the validation set weighted accuracy of classification was estimated. These results are presented in Table 8.

Table 8. Weighted accuracy values based on set B validation structures for 50 generated LR-based models.

Model ID	Weighted accuracy [%]	Model ID	Weighted accuracy [%]
1	91.25	**26**	91.87
2	92.27	**27**	92.20
3	90.32	**28**	89.57
4	92.99	**29**	90.81
5	90.35	**30**	90.43
6	91.31	**31**	90.63
7	90.24	**32**	91.65
8	91.51	**33**	89.37
9	92.61	**34**	91.22
10	93.59	**35**	89.95
11	90.13	**36**	89.88
12	91.32	**37**	91.95
13	91.09	**38**	90.43
14	91.99	**39**	91.38
15	91.24	**40**	91.22
16	90.52	**41**	91.60
17	91.54	**42**	90.91
18	91.61	**43**	91.12
19	92.31	**44**	91.40
20	93.58	**45**	93.43
21	91.57	**46**	91.13
22	89.97	**47**	92.69
23	92.43	**48**	90.82
24	91.04	**49**	91.91
25	91.91	**50**	90.24

The highest obtained value of weighted accuracy was equal to 93.59%, while the lowest was equal to 89.37%. The median and mean values are 91.28% and 91.33%, respectively.

After the lists of selected features were generated, the subsequently occurring genes were collected into one cumulative list along with the weighted accuracy values of the validation structures among 50 generated models. The *GeneRank* measure (16) feature selection filter method was applied to assign the proper feature ranks.

$$GeneRank_x = \sum_{j=1}^{N} \frac{accuracy_j \times (k - i + 1)}{k} \quad (16)$$

Where:
x is the feature indicator
N is the model index
k is the number of features in the most extended model
i is the features' position in the *j*th model

Described importance measure was applied for each feature that occurred at least once in all 50 generated models. *GeneRank* focuses not only on the weighted classification accuracy of the validation set for the model from which the feature is derived but also on the number of features that occurred in the longest-generated model. The last aspect important from the point of view of the introduced measure of genes' importance is the position of the feature in a given model. If a specific gene was present in multiple models, its calculated performances are summed up. After each gene's estimates, the metric values were normalized to the 0-1 range to make all available features comparable. To determine the appropriate cut-off point for the number of informative genes, a feature distribution plot was created along with the assigned values of the normalized *GeneRank* metric. The threshold value for the correct number of features was determined based on the *GeneRank* metric, where the significant difference between the following values was imperceptible. Hence, in Figure 17, presenting the distribution of the metrics' values for individual genes, the threshold was marked with a red line.

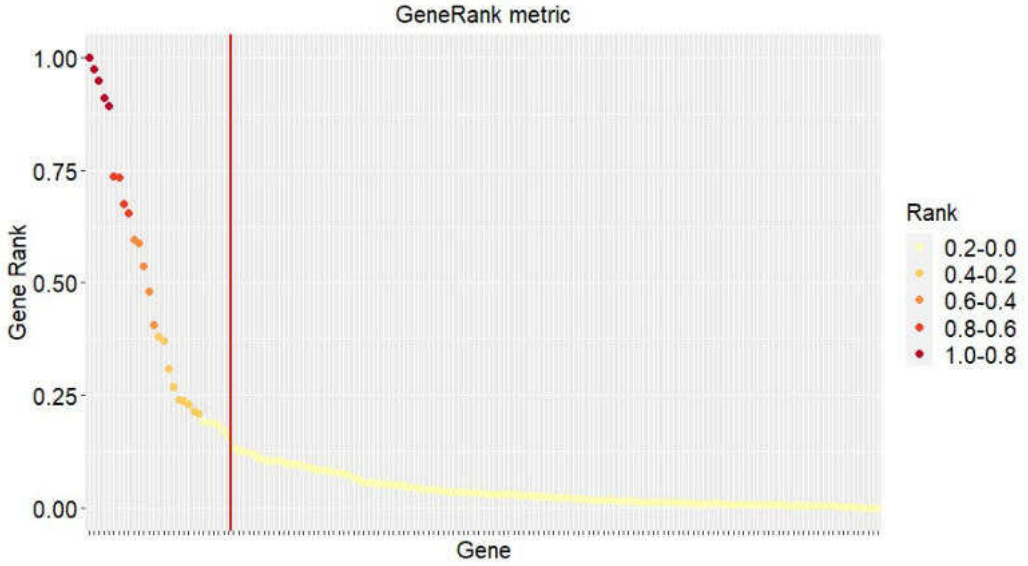


Figure 17. List of importance-sorted features with marked cut-off point for the number of informative genes.

Features on the right of the designated threshold value create a visible continuous line without any jumps between individual values. These genes do not significantly impact the recognition of the genetic profile of ex vivo irradiated cells. They, therefore, are not important from the point of view of the general cells' classification problem. The complete list of sorted features included 159 genes, while after the appropriate cut-off value was determined, 29 genes were left for further analysis of irradiated cells' genetic profile recognition. These were considered informative in terms of the ability to recognize irradiated cells. Chosen 29 genes and their assigned *GeneRank* values are presented in Table 9.

Table 9. Complete list of genes selected for the final model building part.

Gene	*GeneRank*	*Gene*	*GeneRank*
BAX	1.0000	TYMS	0.3715
RPS19P1	0.9743	CD40	0.3087
RPL23AP42	0.9482	TMEM97	0.2678
RPS27L	0.9092	RUNX3	0.2408
DDB2	0.8907	GZMH	0.2388
TNFSF8	0.7362	MYC	0.2313
CCNG1	0.7336	CXCL9	0.2141
STAT5A	0.6756	IL15	0.2098
LCK	0.6556	FYB	0.1923
TNFRSF10B	0.5949	MCM2	0.1905
AQP9	0.5871	FLT3	0.1895
CD3D	0.5361	LAT	0.1856
PHPT1	0.4805	TRIB2	0.1713
AEN	0.4075	GAPDH	0.1592
LAMP3	0.3801		

After features selection and determination of genes' significance in the problem of distinguishing between control and irradiated cells, there were estimated model parameter values utilizing the implemented algorithm. The calculations were based on the model structure of the B set, and estimated parameter values are presented in Table 10.

Table 10. Estimated parameter values for the final model.

Intercept	**BAX**	**RPS19P1**	**RPL23AP42**	**RPS27L**	**DDB2**	**TNFSF8**	**CCNG1**
-2.47	0.79	0.25	-0.21	0.74	1.27	0.71	0.48
STAT5A	**LCK**	**TNFRSF10B**	**AQP9**	**CD3D**	**PHPT1**	**AEN**	**LAMP3**
-0.6**3**	-0.29	0.85	0.39	-0.21	0.34	0.90	-0.17
TYMS	**CD40**	**TMEM97**	**RUNX3**	**GZMH**	**MYC**	**CXCL9**	**IL15**
-0.69	-0.33	-0.47	-0.27	-0.20	-0.19	-0.06	-0.40
FYB	**MCM2**	**FLT3**	**LAT**	**TRIB2**	**GAPDH**		
-0.3**1**	0.29	-0.26	-0.25	-0.47	-0.14		

The last step necessary to build a fully functioning, tuned model was to select the appropriate probability threshold value for irradiated cells classification. This step will arm the built model with an adjusted probability level, allowing even more efficient cell classification. For this purpose, a classification threshold inspection step was performed based on the set B test structure, utilizing the Youden index (17).

$$J = Sensitivity + Specificity - 1 \quad (17)$$

It is a valuable metric, especially if the goal is to select the optimal cut-off classification threshold for a balanced study of sensitivity and specificity measures. The Youden index is an indicator specified for all Receiver Operating Characteristic (ROC) points. The maximum value of the indicator may be used as a criterion for selecting the probability value for classification purposes. The index is often

represented graphically as the height above the diagonal line. The determined ROC curve with the estimated 0.7047 value of the new classification probability value is presented in Figure 18.

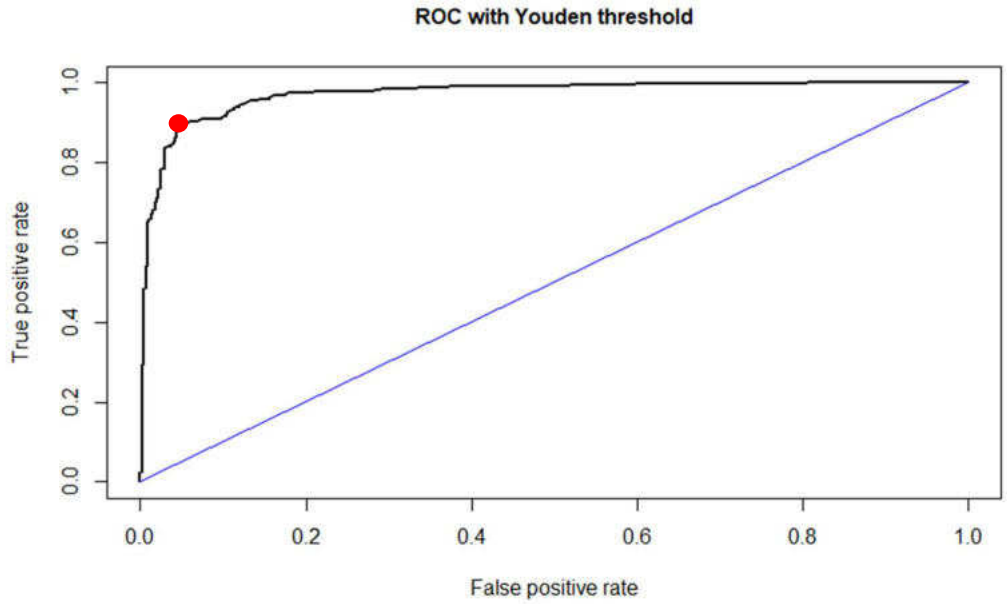


Figure 18. ROC with Youden classification probability value marked.

A testing procedure was performed using the set B test data structure to verify whether a significant change occurred after utilizing the new classification probability threshold value. To compare the results, several qualitative metrics were introduced: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), precision, recall, specificity, and weighted classification accuracy. All these measures are described in detail in Table 11 and Table 12.

Table 11. Confusion matrix details.

		True class	
		Positive	Negative
Predicted class	Positive	TP	FP
	Negative	FN	TN

Table 12. Introduced qualitative metrics.

Metric name	**Formula**
Precision	TP/(TP+FP)
Recall (Sensitivity)	TP/(TP+FN)
Specificity	TN/(TN+FP)
Weighted accuracy	(Recall + Specificity)/2

Precision is the ratio of observations correctly classified as positive to all marked positives. Sensitivity (recall), on the other hand, is the ratio of observations correctly classified as a positive group concerning all positive cases in the actual data set. The explanation for specificity is almost the same as sensitivity, but on the contrary, it examines the negative group. Additionally, the $F1_{score}$ (18) measure was introduced, enabling the comparison of different models' quality metrics. The $F1_{score}$ metric takes into account both FPs and FNs. It is, therefore, a much more valuable metric than the pure classification quality value.

$$F1 = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \quad (18)$$

To verify the effectiveness of the irradiated cells' new classification probability value, the results for the default and new classification probability thresholds are presented in Table 13. This analysis was carried out based on the set B test structure.

Table 13. Classification quality metrics comparison for two classification probability threshold values.

	Classification threshold	
Quality metric	**Fixed 0.5000**	**Youden 0.7047**
TP	361	351
TN	404	429
FP	48	23
FN	30	40
Precision	0.8826	0.9385
Sensitivity	0.9233	0.8977
Specificity	0.8938	0.9491
Weighted accuracy	0.9075	0.9253
$F1_{score}$	0.9025	0.9176
Number of cells	843	843
Number of correctly classified cells	765	780
Number of incorrectly classified cells	78	63
Incorrectly classified cells [%]	9.25	7.47

The estimated $F1_{score}$ metric values indicate the advantage of classification quality for the changed probability classification threshold. Utilizing the new threshold value, it was possible to classify correctly 15 cells more compared to the predetermined threshold. The operations related to the feature selection and the refinement of the final model made it possible to obtain a weighted classification quality at a very high and satisfactory level, amounting to over 92%.

As the final verification of the model's accuracy, the independent testing procedure was performed based on the set A data. It was necessary to perform one more additional checking step to determine factors precluding the possibility of direct testing on this dataset. There was investigated if the batch effect is present among sets distribution in the two-dimensional UMAP space. Before UMAP projection, the dimensionality of both data sets (set A and set B) was limited using the PCA method to reduce computational costs. There were analyzed all observations from both ex vivo datasets. Moreover, there were considered raw datasets without preliminary quality control steps. As a result of the PCA, the analyzed dimensionality was reduced from 452 to only 18 PCA components, explaining 96.74% of the variance in the pooled dataset. The data was next transformed into a two-dimensional

UMAP space and projected using isolines showing the distribution of the data cloud of both datasets. The results are presented in Figure 19.

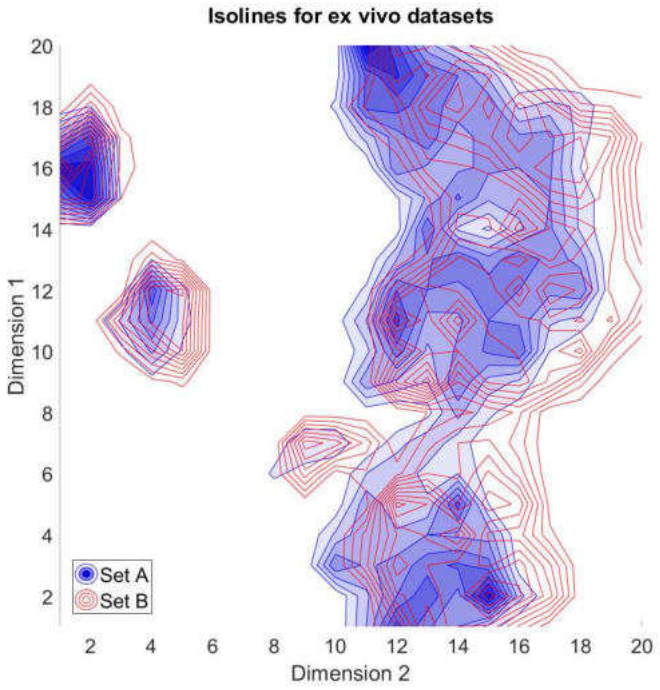


Figure 19. Isolines for ex vivo datasets distribution.

The distributions of both ex vivo datasets are consistent, as they occupy very similar areas in the two-dimensional space. What is more, the presented distribution overlap between both datasets. The testing procedure can be performed on the raw set A data without additional data manipulation and processing. The results of the independent testing procedure are presented in Table 14.

Table 14. Classification quality metric values based on the independent test set.

Quality metric name	Quality metric value
TP	1017
TN	1485
FP	99
FN	122
Precision	0.9113
Sensitivity	0.8929
Specificity	0.9375
Weighted accuracy	0.9188
$F1_{score}$	0.9020
Number of cells	2723
Number of correctly classified cells	2502
Number of incorrectly classified cells	221
Incorrectly classified cells [%]	8.12

Results of the testing procedure based on an independent dataset yielded very satisfactory results. These are reflected in the high value of the weighted classification accuracy, exceeding 91%. There were also achieved high specificity and precision values of above 0.90. This means that the constructed classifier correctly recognizes both irradiated and control cells.

6.2 White blood cell subpopulations recognition

After a detailed analysis of individual genes included in the recognized genetic profile of irradiated cells, it turned out that not all genes are used to distinguish irradiated cells from control cells. Some genes in the recognized profile have other functions, including recognition of specific cell subtypes. Table 15 is based on supplementary material *BD Rhapsody Immune Response Panel Hs*, provided with the single-cell sequencing data. It was found that as many as 9 out of 29 genes of the irradiated cells' genetic profile are responsible for distinguishing individual cell subpopulations. They constitute an additional burden for the built genetic profile.

Table 15. Genes of the recognized irradiated cells' genetic profile with assigned corresponding functions.

Gene name	Function/process – cell type specificity
BAX	Apoptosis regulator
RPS19P1	-
RPL23AP42	-
RPS27L	-
DDB2	-
TNFSF8	Cytokine
CCNG1	-
STAT5A	Transcription factor
LCK	**Marker gene – T subset**
TNFRSF10B	-
AQP9	Transporter
CD3D	**Marker gene – Pan T**
PHPT1	-
AEN	-
LAMP3	**Marker gene – Dendritic cells**
TYMS	Cell cycle (S phase)
CD40	CD marker
TMEM97	Miscellaneous
RUNX3	**Transcription factor – T subset**
GZMH	**Effector molecule – Cytotoxic T, NK**
MYC	Proliferation marker
CXCL9	Chemokine
IL15	Interleukin
FYB	**Miscellaneous – T subset**
MCM2	Cell cycle (S phase)
FLT3	**CD marker, kinase – Dendritic cells**
LAT	**Miscellaneous – T subset**
TRIB2	**Kinase - T**
GAPDH	Metabolism

These genes are not directly accountable for determining cells due to the analyzed external factor, so they do not fully correspond to the aim of this doctoral dissertation which focuses primarily on radiation response genes. Undoubtedly, however, they play an important role in distinguishing irradiated and control cells. Observations from different subpopulations of white blood cells are characterized by differences in the strength of response to ionizing radiation, as described in the introduction to the doctoral dissertation *White Blood Cell subpopulations*. This analysis allowed us to discover a critical correcting factor in the problem of the irradiated cells' genetic profile recognition - the internal heterogeneity of the white blood cells dataset influencing the final irradiated cells' genetic signature.

Due to the purpose of this dissertation, which is constructing a classifier for control and irradiated cell recognition, the detected heterogeneity of the dataset must be carefully analyzed. The aim is to eliminate internal differences between cells that do not result from the influence of the radiation factor but influence the behavior of gene profiles of these cells. To determine the exact cause of the variation in collections, procedures were carried out to isolate individual subpopulations of white blood cells. The cell subpopulations were analyzed separately for the model structures of both technical repetitions and the set B test structure. The problem of white blood cell subpopulation recognition has been divided into the following stages: feature selection, cell cluster recognition utilizing the HDBSCAN tool, and white blood cell subpopulations recognition using marker genes characteristic of the expected cell subpopulations (*BD Rhapsody Immune Response Panel Hs*).

6.2.1 Feature selection

Before the relevant part of WBCs subpopulation recognition, the feature selection procedure was performed. It is primarily aimed at reducing the dimensionality of the datasets. The goal is to select the genes set that show the most significant variability across the control and irradiated cell count values. For this purpose, the coefficient of variation (CV) measure was used, according to (19).

$$CV_{gene} = \frac{standard\ deviation_{gene}}{mean_{gene}} \times 100\% \qquad (19)$$

In the analyzed data, different genes are characterized by varying counts. The CV measure benefits the variables with various ranges of values comparison procedures. The lower the estimated CV value, the more stable the feature is, which means the lower CV values are dedicated to fewer variable features over the analyzed set. The dimensionality reduction using the described measure of the variability is based on 396 genes from set A and 406 genes from set B (available in the data sets after QC procedures were performed). The GMM was utilized with the number of GMM components determined by the BIC metric to determine the appropriate threshold value for the number of features. A reasonable threshold value was estimated based on the CV and the GMM approaches. Among both analyzed datasets, this value was determined by rejecting the one component that contained features with the lowest variability. The feature selection results for both model structures and the set B test structure are presented in Figure 20, Figure 21, and Figure 22, respectively.

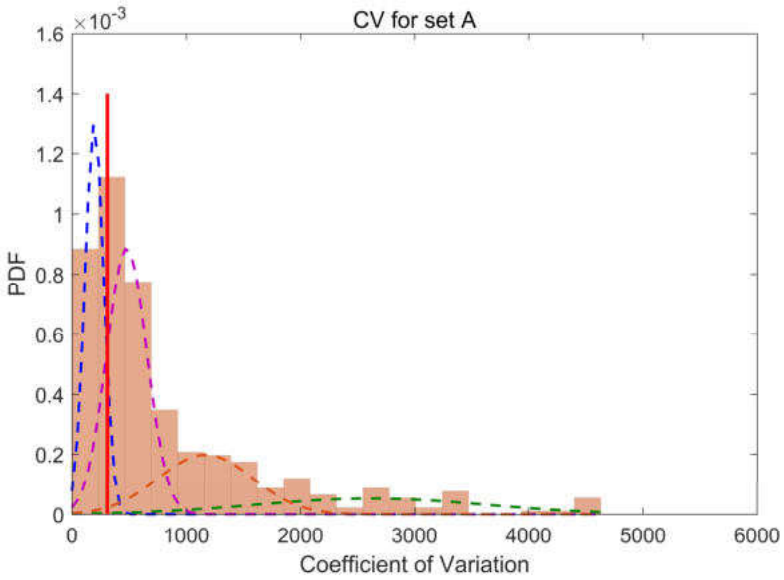


Figure 20. GMM components for CV distribution with the number of chosen features threshold marked based on set A model structure.

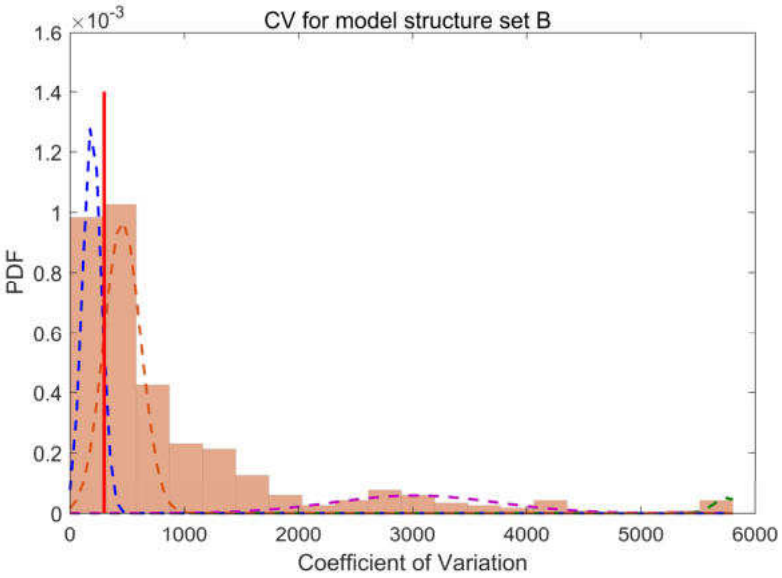


Figure 21. GMM components for CV distribution with the number of chosen features threshold marked based on set B model structure.

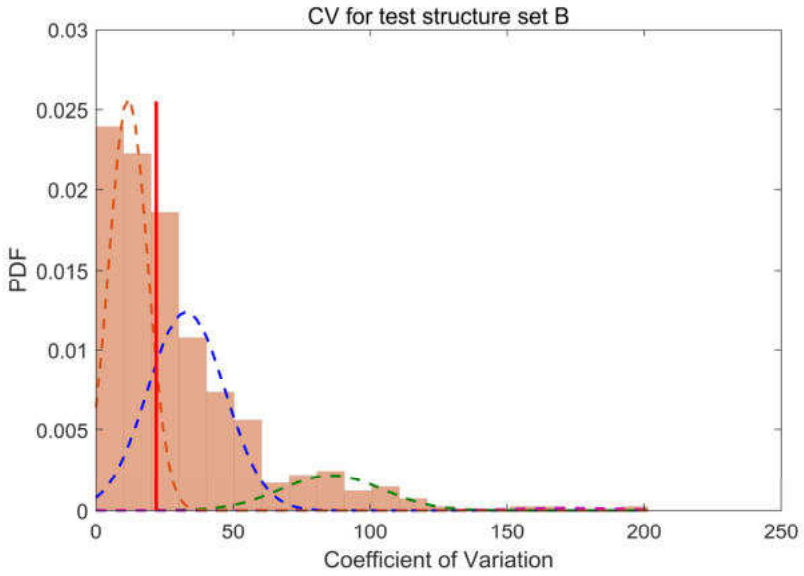


Figure 22. GMM components for CV distribution with the number of chosen features threshold marked based on set B test structure.

The summary of performed dimensionality reduction procedure with the estimated CV threshold values and the number of selected genes and cells are described in Table 16. In addition, for both model structures for ex vivo data, 46 and 14 cells were detected, respectively, for set A and set B data, having non-zero counts in less than three genes available after dimensionality reduction using CV. Described cells were marked suspicious and removed from further analysis based on the HDBSCAN tool. In the next step of the subpopulation recognition part, these cells will be restored to the model datasets according to the proposed procedures for including sub-types of cells into individual clusters.

Table 16. Summary of feature selection procedures for the WBCs subpopulation recognition, detailing the analyzed data structures.

Data	CV threshold	Genes	Control cells	Irradiated cells	Total cells
Set A model	310	273	1234	866	2100
Set B model	300	286	1796	1561	3357
Set B test	22	197	452	391	843

6.2.2 HDBSCAN cluster analysis

The HDBSCAN tool was utilized to divide cells into clusters corresponding to their variability across the spatially reduced datasets. The cluster separation with the tool consisted mainly of selecting an appropriate starting parameter set and properly dividing the data set into clusters. The influence of 3 out of 4 properties listed in Table 17 was analyzed to select the appropriate set of parameters.

Table 17. Set of starting HDBSCAN tool parameters with their interpretation.

Parameter	Interpretation
min_cluster size *(mcs)*	the smallest group size that is considered to be a cluster
min_samples *(ms)*	the measure how conservative the clustering algorithm is – the larger the value, the more conservative clustering (the larger the value, the more points will be declared as noise, and clusters will be restricted to more dense areas)
cluster_selection_epsilon *(cse)*	helps to merge clusters in regions determined by parameter value - if even groups of few points might be of interest to us, we can change this parameter
metric	euclidean

A total of 40 different sets of starting parameters were tested based on the original count values considering only the control cell samples. The analysis was performed separately for each of the three data sets: set A model structure, set B model structure, and set B test structure. The effect size omega-squared measure was used to determine the parameter set that best separates the analyzed control cells in terms of their differentiation, according to (20).

$$\omega^2 = \frac{SS_{between} - df_{between} \times MS_{within}}{SS_{total} + MS_{within}} \quad (20)$$

Where:
$SS_{between}$ is the between-group variation
$df_{between}$ is the between group degrees of freedom
MS_{within} is the mean square within groups
and SS_{total} is the total variation

The estimates of the omega-squared measure took into account all genes available after the features selection procedure with the use of CV and clusters generated by the HDBSCAN tool, excluding groups containing unassigned cells. The complete set of tested parameters and the results of the analysis performed for individual data sets are included in Table 18. The selection of an appropriate set of parameters was based primarily on the value of the effect size measure and the content of unclassified cells (*Uncls. cells*). The selected sets of parameters for individually analyzed datasets are also marked in Table 18 – for the set A and set B model structures, the set of parameters marked with the number 38 was chosen. The parameters marked with the number 3 were selected for the set B test structure.

Table 18. Set of tested HDBSCAN starting parameters.

				Set A model structure			Set B model structure			Set B test structure		
Param. version	***mcs***	***ms***	***cse***	No. clusters	Uncls. cells [%]	ω^2	No. clusters	Uncls. cells [%]	ω^2	No. clusters	Uncls. cells [%]	ω^2
1	10	2	0.100	2	13.13	0.75	6	10.80	0.97	2	1.99	0.75
2	15	2	0.100	2	13.13	0.75	3	12.64	0.72	2	1.99	0.75
3	5	2	0.100	7	11.43	0.85	13	8.91	0.99	3	0.66	0.86
4	5	5	0.100	5	14.26	0.73	5	12.47	0.98	2	3.32	0.65
5	5	10	0.100	2	16.05	0.38	2	14.20	0.46	0	100.00	0.00
6	5	2	0.300	7	11.43	0.85	13	8.91	0.99	3	0.66	0.86
7	5	2	0.050	7	11.43	0.85	13	8.91	0.99	3	0.66	0.86
8	5	2	0.001	7	11.43	0.85	13	8.91	0.99	3	0.66	0.86
9	20	2	0.100	2	13.13	0.75	2	12.53	0.50	0	100.00	0.00
10	20	5	0.100	2	14.75	0.67	2	13.81	0.61	0	100.00	0.00
11	20	6	0.100	2	15.48	0.57	3	55.40	0.11	0	100.00	0.00
12	20	10	0.100	3	61.91	0.33	2	48.55	0.05	0	100.00	0.00
13	20	8	0.100	2	15.56	0.56	7	56.57	0.34	0	100.00	0.00
14	15	8	0.100	2	15.56	0.56	2	42.87	0.05	0	100.00	0.00
15	15	9	0.100	3	50.24	0.18	2	13.86	0.57	0	100.00	0.00
16	15	5	0.100	2	14.75	0.67	2	13.81	0.61	0	100.00	0.00
17	10	8	0.100	2	15.56	0.56	2	14.31	0.45	0	100.00	0.00
18	10	8	0.200	2	15.56	0.56	2	14.31	0.45	0	100.00	0.00
19	10	8	0.300	2	15.56	0.56	2	14.31	0.45	0	100.00	0.00
20	10	8	0.400	2	15.56	0.56	2	14.31	0.45	0	100.00	0.00
21	10	8	0.600	2	15.56	0.56	2	14.31	0.45	0	100.00	0.00
22	5	8	0.200	2	15.56	0.56	4	13.47	0.60	2	38.72	0.06
23	5	8	0.050	2	15.56	0.56	4	13.47	0.60	2	38.72	0.06
24	2	8	0.200	6	15.32	0.90	8	12.81	0.97	2	4.42	0.41
25	10	11	0.200	2	15.96	0.49	2	14.20	0.46	0	100.00	0.00
26	30	35	0.200	2	59.81	0.14	3	63.70	0.11	0	100.00	0.00
27	30	10	0.200	3	61.91	0.33	5	60.75	0.25	0	100.00	0.00
28	6	8	0.100	2	15.56	0.56	4	13.47	0.60	2	38.72	0.06
29	5	2	0.000	7	11.43	0.8	13	8.91	0.99	3	0.66	0.86
30	5	2	0.020	7	11.43	0.85	13	8.91	0.99	3	0.66	0.86
31	10	4	0.005	2	14.10	0.72	2	13.70	0.60	2	2.65	0.73
32	8	4	0.005	2	14.10	0.72	4	12.69	0.60	2	2.65	0.73
33	5	4	0.010	5	13.53	0.79	7	11.80	0.98	3	1.55	0.86
34	5	4	0.000	5	13.53	0.79	7	11.80	0.98	3	1.55	0.86
35	5	6	0.000	3	15.07	0.66	5	12.75	0.75	2	3.76	0.56
36	5	1	0.000	7	11.43	0.85	13	8.91	0.99	3	0.66	0.86
37	10	1	0.000	2	13.13	0.75	6	10.80	0.97	2	1.99	0.75
38	3	1	0.000	13	10.05	0.97	24	7.80	0.99	4	0.88	0.86
39	4	1	0.000	9	10.78	0.95	16	8.24	0.99	3	0.66	0.86
40	6	1	0.000	5	12.24	0.81	11	9.47	0.99	3	0.66	0.86

Next, the control and irradiated cells were subjected to the clustering procedure. The HDBSCAN tool starting parameter sets selected from the control samples were applied to the control and the irradiated sample for the specified dataset. An extension of this tool was also used to assign cells of unknown origin to the closest, based on the estimated probability value, clusters. As a result of the ex vivo data analysis with the HDBSCAN tool, the cells were divided into clusters. These clusters are assumed to correspond to the intrinsic heterogeneity of the cells, which was detected in the ex vivo datasets

visualization process. The result of the operation of this tool, for each analyzed cell, is the vector of probabilities of this cell belonging to all of the detected by HDBSCAN tool clusters. The assignment was utilized for the cluster with the highest probability of belonging value to determine the membership of individual cells. In this way, all analyzed cells were described by the unique cluster ID showing their internal variability. Table 19 shows the number of cell clusters detected for the control and irradiated sample, depending on the analyzed dataset.

Table 19. The number of detected clusters for ex vivo data sets.

Data set	Cell type	Number of HDBSCAN clusters
Set A model structure	Control	13
	Irradiated	11
Set B model structure	Control	24
	Irradiated	23
Set B test structure	Control	3
	Irradiated	3

This step also restored weak-count cells previously removed from the analysis at the features selection step using the CV metric for both model structures of the ex vivo experiment (46 cells for set A and 14 cells for set B data). The centroids-based distance metric was used for cell clusters detected by the HDBSCAN tool. This procedure was applied separately to the control and irradiated cells. To achieve the best possible accuracy in restoring weak-count cells, the full dimensionality of the data was restored that was available after QC procedures. In the case of the set A and set B datasets, the dimensionality of respectively 396 and 406 features was restored. The centroid values were determined as multidimensional structures for each available feature in the next step. Each dimension of a centroid is related to the specified feature. Weak-count cells were restored to the dataset severally, and Spearman's correlation measure was calculated for each available cluster. The selected cell was assigned to the cluster for which the highest estimated correlation value was achieved. The weak-count cell recovery procedure was performed on the original scaled data. Thus, the data was returned to their full dimensionality, containing the full range of cells available after QC procedures. In addition to restoring full dimensionality, each cell was assigned to a specific data heterogeneity cluster.

6.2.3 WBC subpopulations recognition

The stage allowing the detection of specific subpopulations of white blood cells was performed using the information about marker genes characteristic for particular subpopulations of WBCs. Table 20 presents a list of the marker genes used, with the assignment of the appropriate subpopulation of WBCs.

Table 20. List of marker genes with an assignment of the corresponding subpopulation of white blood cells.

WBCs subpopulation	Marker gene
Monocytes	FCGR3A, S100A9, CD14
Dendritic cells	LAMP3, CD1C
T cells	CD3G, CD4, LCK, LEF1, SELL, FOXP3, CD8A, CD8B
B cells	CD79A, CD79B, TCL1A, MS4A1
Granulocytes	PI3
Basophil/Eosinophil	CLC
NK cells	CST7, CTSW, NKG7
Alpha/Beta T cells	TRAC, TRBC2
Gamma/Delta T cells	TRDC

To identify the relevant structures of the WBC subpopulations, including even very low-represented cell types, boxplots were generated for each cell cluster and each subpopulation marker gene, showing the counts distribution. Based on generated boxplots, individual clusters belonging to the appropriate cell subpopulation were decided separately for the control and irradiated cells. The clusters marked as the same subpopulation were joined, creating larger structures corresponding to the internal data heterogeneity. Examples of boxplots for the three analyzed ex vivo data structures for the CD14 monocyte subpopulation marker gene are presented in Figure 23, Figure 24, and Figure 25. The remaining boxplots can be found in the section *Additional materials* in the subsection *Recognition of cell subpopulations based on ex vivo experiments.* Control cell clusters (C) are marked in goldenrod, while irradiated cell clusters (R) are kept in navy blue.

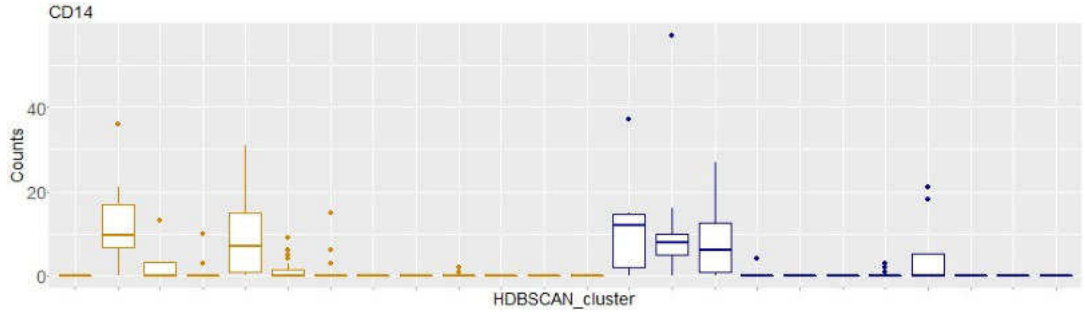


Figure 23. Boxplot of count distribution for the CD14 monocyte marker gene for set A model structure data.

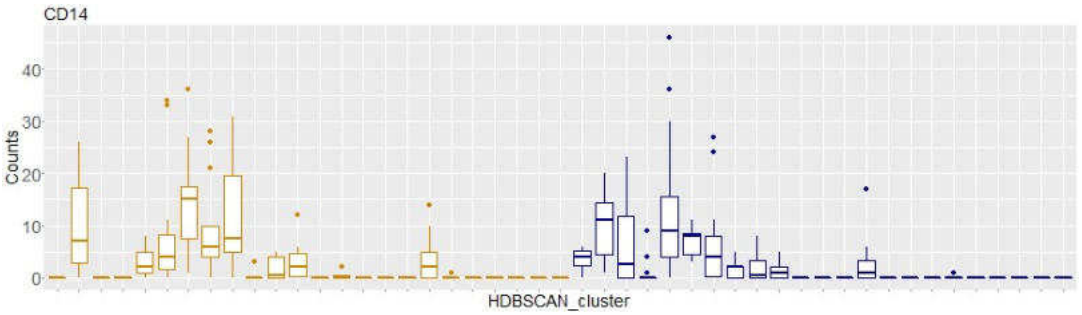


Figure 24. Boxplot of count distribution for the CD14 monocyte marker gene for set B model structure data.

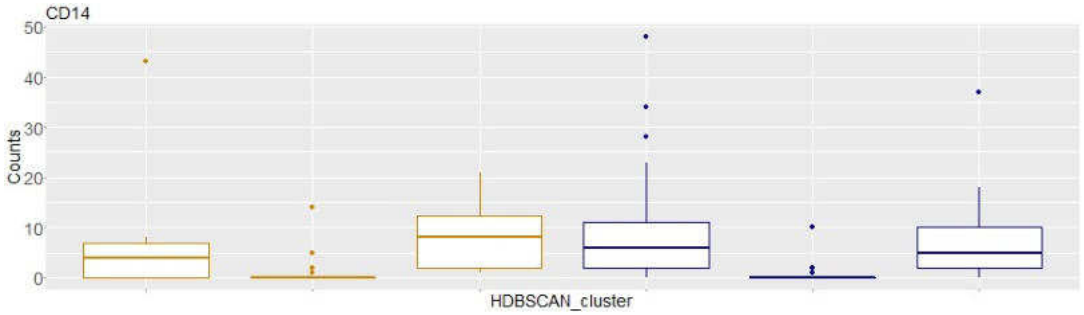


Figure 25. Boxplot of count distribution for the CD14 monocyte marker gene for set B test structure data.

The detected WBC subpopulations were visualized in the next step using the UMAP tool. The designated subpopulations were marked with different colors to test whether the structure of clusters, defined using unsupervised learning methods, corresponds to the recognized cell subpopulations. After the analysis of the resulting UMAP projection plots, it was noticed that in the case of both set A and set B model data structures, there exists a small fraction of cells that are mixed with a subpopulation other than assigned. Figure 26 and Figure 27 show the described phenomenon successively for the set A and B model data structures. In the case of the model structures of the set A data, 19 cells (14 control and 5 irradiated), and the set B data, 19 such cells (9 control and 10 irradiated) were marked suspect.

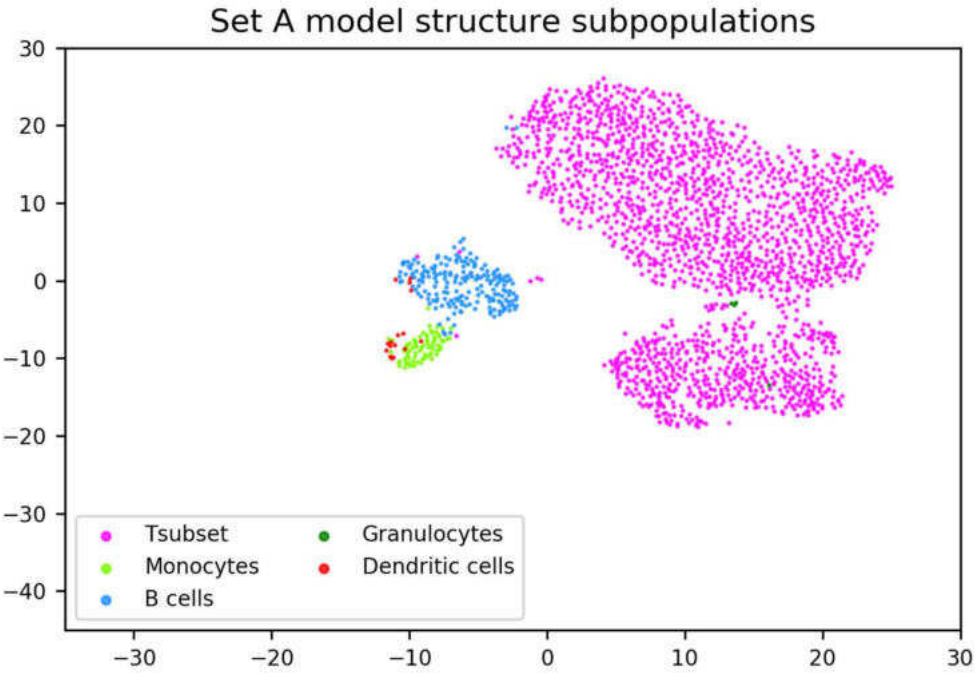


Figure 26. UMAP visualization for detected WBC subpopulations for set A model structure.

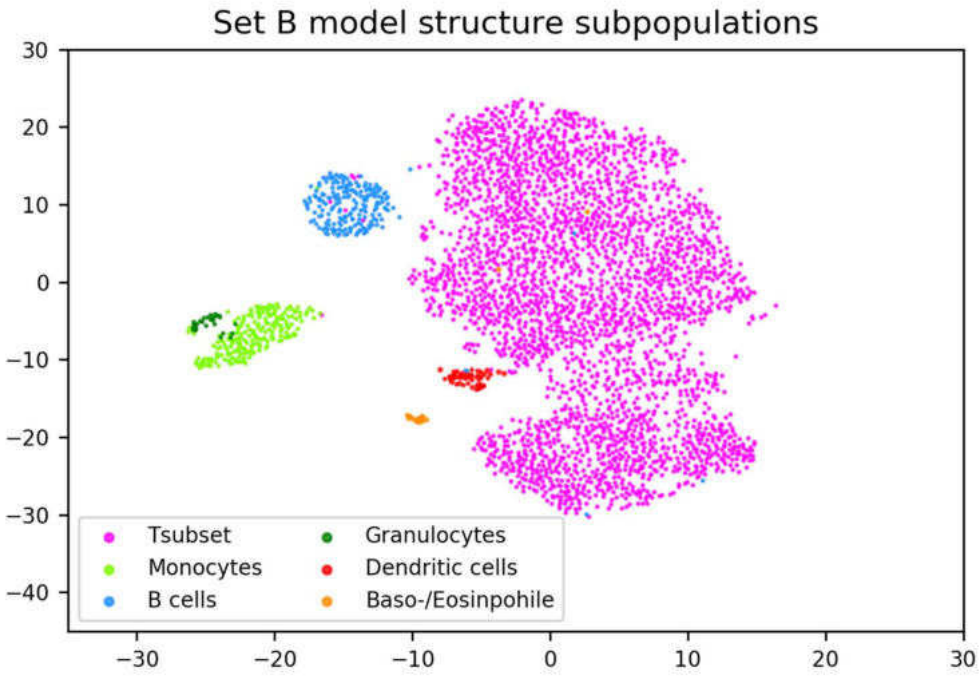


Figure 27. UMAP visualization for detected WBC subpopulations for the set B model structure.

Suspicious cells were subjected to a detailed analysis to match them to the genetic profiles of the remaining correctly classified subtypes of white blood cells. This procedure consisted of redetermining the boxplots for marker genes. However, the distribution for not all cells classified as a given subtype is plotted in this case. In this case, there are marked only those cells for which there is no doubt their affiliation to a specific subpopulation type is correct. The suspicious cell counts were plotted directly on the distributions of individual cell subpopulations. If the suspect cell shows increased counts for the marker gene of the subpopulation they do not belong, it means that they have been incorrectly assigned, and it is necessary to change their affiliation. Figure 28 and Figure 29 show, for the model structure of the set A and set B data, an exemplary procedure for the CD14 gene, which is a marker gene for the monocyte subpopulation. Other generated boxplots can be found in *Additional materials* in the subsection *Recognition of suspicious cells affiliation.*

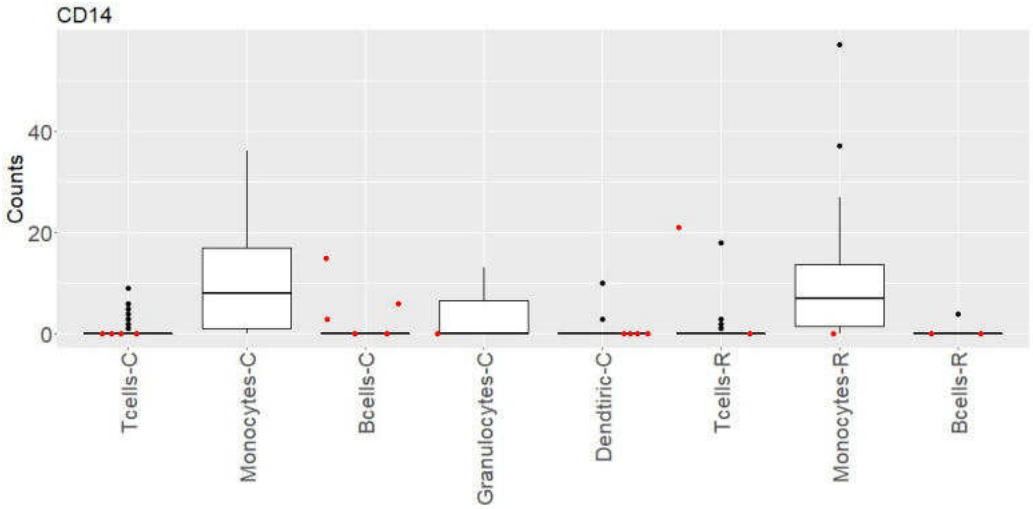


Figure 28. Analysis of the suspect cells affiliation based on the set A model structure data.

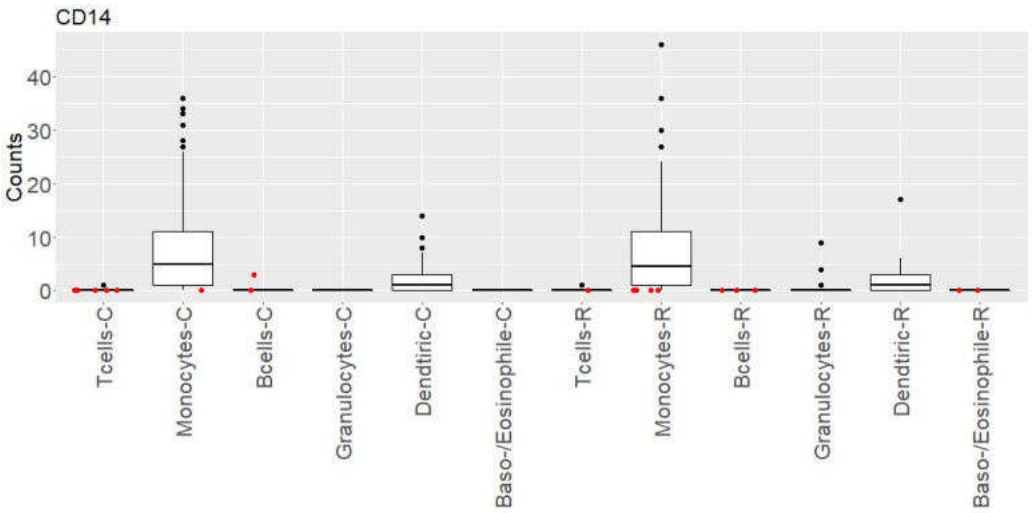


Figure 29. Analysis of the suspect cells affiliation based on the set B model structure data.

Analyzing the boxplots in Figure 28 and the T-cell subpopulation for irradiated cells (*Tcells-R*), it can be seen that one of the red-marked cells shows increased counts for the CD14 monocyte marker gene. Therefore, this cell should be reclassified from a T-cell to a monocyte subpopulation. Described procedure was performed for all suspect cells from the set A and set B model structures data.

After thoroughly analyzing suspicious cells for both model sets, the heterogeneity structures of the analyzed data were presented utilizing 2-dimensional UMAP projection. The final UMAP plots with marked identified cellular subpopulations are shown in Figure 30, Figure 31, and Figure 32 for the model structures of set A and set B and for the set B test structure, respectively.

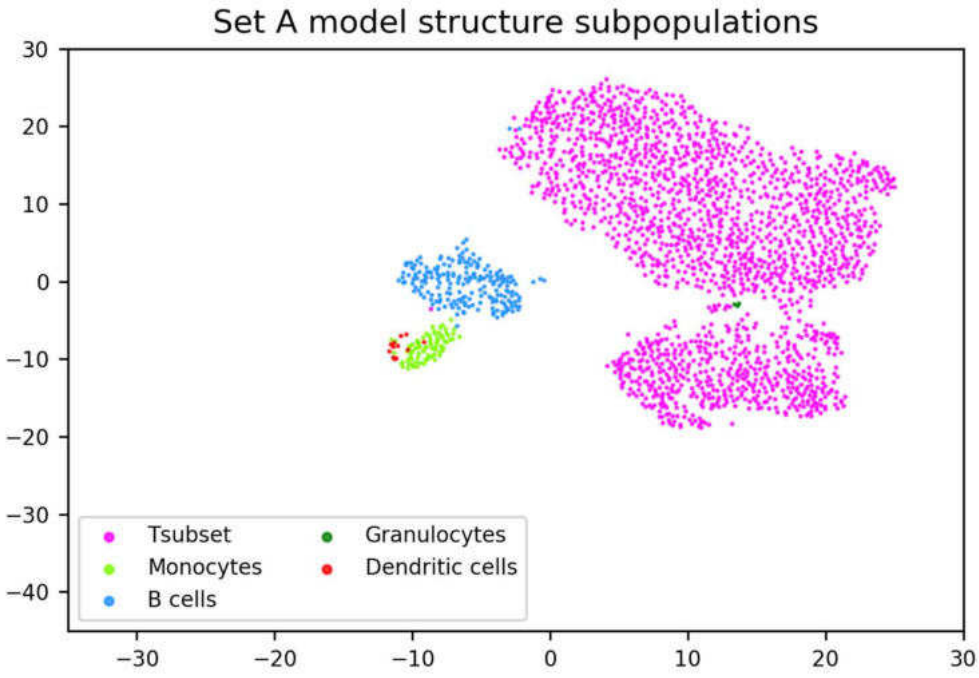


Figure 30. UMAP visualization for final WBC subpopulations based on set A model structure data.

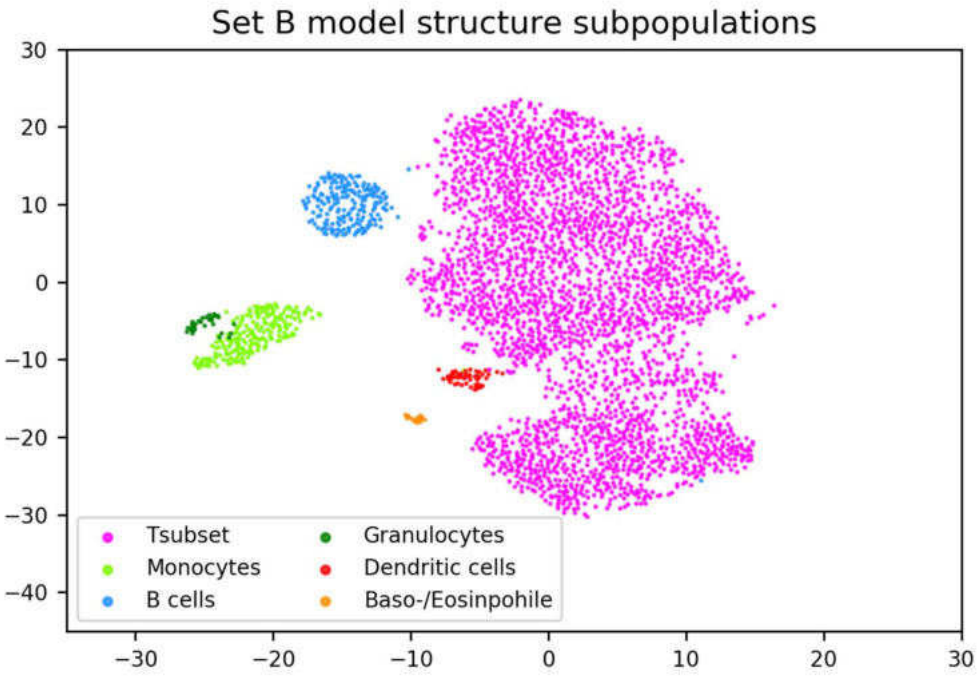


Figure 31. UMAP visualization for final WBC subpopulations based on set B model structure data.

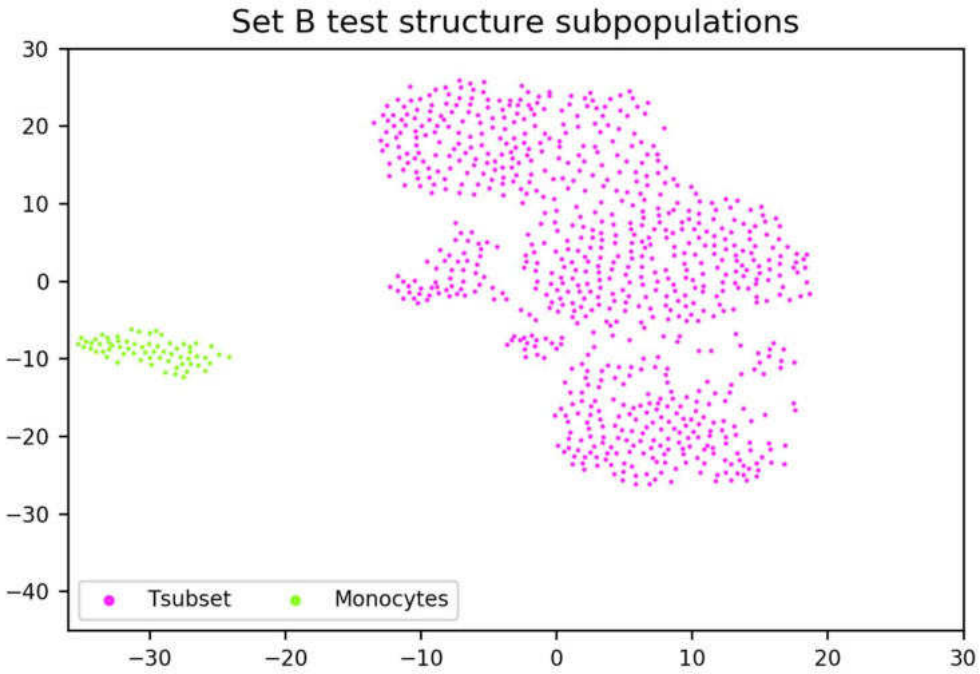


Figure 32. UMAP visualization for final WBC subpopulations based on set B test structure data.

As a result of the re-estimation of the individual WBCs subpopulation distribution, it was found that there is a noticeable improvement in cell affiliation based on the UMAP projection. A summary of detected white blood cell subpopulations based on the ex vivo data is presented in Table 21. It contains the numerical and percentage share of cells of a specific cell subpopulation concerning the analyzed datasets.

Table 21. Summary of the white blood cell subpopulation recognition.

WBCs subpopulation	Cell type	Set A model structure	Set B model structure	Set B test structure
T cells	Control	83.88% (1051)	85.56% (1540)	94.91% (429)
	Irradiated	89.70% (801)	84.66% (1330)	89.26% (349)
Monocytes	Control	3.67% (46)	6.28% (113)	5.09% (23)
	Irradiated	3.47% (31)	5.60% (88)	10.74% (42)
B cells	Control	11.25% (141)	5.56% (100)	not detected
	Irradiated	6.83% (61)	6.36% (100)	not detected
Granulocytes	Control	0.24% (3)	1.44% (26)	not detected
	Irradiated	not detected	1.59% (25)	not detected
Dendritic cells	Control	0.96% (12)	0.61% (11)	not detected
	Irradiated	not detected	1.34% (21)	not detected
Basophile/ Eosinophile	Control	not detected	0.55% (10)	not detected
	Irradiated	not detected	0.45% (7)	not detected

Recognized WBC subpopulations overlap very well with the structures generated by the unsupervised approach to data clustering using the UMAP tool. The shares of individual subpopulations about a given data set presented in Table 21 indicate a significant advantage of the content of T-cells in all analyzed sets of cells. Moreover, no essential differences can be observed in the percentage of cells of a given subpopulation concerning the control and irradiated groups. Thanks to carefully conducted analysis, it was also possible to detect very few subpopulations, such as granulocytes, dendritic cells, or basophils/eosinophils. This indicates a high sensitivity of the performed analysis aimed at cell subpopulation recognition. A very in-depth analysis of the cells marked as suspicious was focused on a relatively small group of cells. Therefore, it can be pondered if the lack of this part of the analysis will significantly disturb the conclusion about the recognition of cell subpopulations. The answer to this hypothesis is not easy, but it is essential to pay attention to the purpose of the study. The aim was to recognize as much heterogeneity as possible in the datasets. Such a thorough study of the structures hidden inside the data allowed for accurate recognition of the cause of cell variability detected using unsupervised machine learning techniques. Moreover, the detected phenomenon was confirmed by several visualizations, which proved the presence of variability due to the WBC subpopulation's existence. Omitting the additional step of correcting subpopulations due to suspicious cells would result in lower accuracy in the assignment of individual cells, thus leaving part of the variability in a way impossible to control later. Building a self-learning classifier to distinguish between control and irradiated cells will undoubtedly benefit from a thorough analysis of cell subpopulations. In the data set prepared this way, the cause of uncontrolled and visible variability within the cells was eliminated. Considering all the aspects mentioned above of the WBC subpopulation recognition procedures, it can be clearly stated that the proposed solution enables, in a very satisfactory way, to isolate cell clusters corresponding to their biological meaning in the case of ex vivo data.

6.3 Irradiated cells' genetic profile recognition based on T-cells subpopulation

Detection of cell subpopulations revealed bases of the heterogeneity visible in 2-dimensional UMAP plots using unsupervised techniques. Utilizing the previously determined genetic profile of irradiated cells and the model built on its basis, it is also possible to decide on the models' effectiveness in classifying observations belonging to selected subpopulations of white blood cells. A summary of classification quality results for individual cell subpopulations based on the independent set A model structure is provided in Table 22. The table shows the three cell subpopulations detected for control and irradiated samples.

Table 22. Classification quality metric values based on the independent test set for three recognized subpopulation types.

Quality metric name	T-cells	Monocytes	B-cells
TP	749	20	54
TN	995	33	106
FP	56	13	35
FN	52	11	7
Precision	0.9304	0.6061	0.6067
Sensitivity	0.9351	0.6452	0.8852
Specificity	0.9467	0.7174	0.7518
Weighted accuracy	0.9417	0.6883	0.7921
$F1_{score}$	0.9328	0.6250	0.7200
Number of cells	1852	77	202
Number of correctly classified cells	1744	53	160
Number of incorrectly classified cells	108	24	42
Incorrectly classified cells [%]	5.83	31.17	20.79

Based on the quality analysis of the classification of individual subpopulations, it was found that only the T-cells subpopulation is correctly classified, maintaining satisfactory results. All values of the analyzed quality metrics are higher for the T-cells subpopulation than for the model based on complete data for all WBC subtypes. This proves the low quality concerning the classification of cell subtypes, constituting a significant minority of the analyzed data set. The weighted classification quality for Monocytes and B-cells is 69% and 79%, respectively, while for the T-cells set, it is as high as 94%. Significantly lower values of this metric, with simultaneous consideration of the $F1_{score}$ measure, mean that the model has learned specific patterns adequate to most T-cells, attaching less importance to the other, less numerous cell subpopulations.

The distribution of counts for the selected set of genes for the detected cell subpopulations was also analyzed. Distributions were drawn using boxplots separating control and irradiated samples. The analysis, therefore, concerned each cell subpopulation and each sample individually. The aim was to determine and confirm the specificity of some genes relating to the data's internal heterogeneity. This analysis was performed based on data from the set B model structure. The selected three genes are shown in the following figures: Figure 33, Figure 34, and Figure 35. The distribution of the remaining genes is presented in the *Additional materials* section in the *Distribution of counts for selected genes by cell subpopulations* subsection.

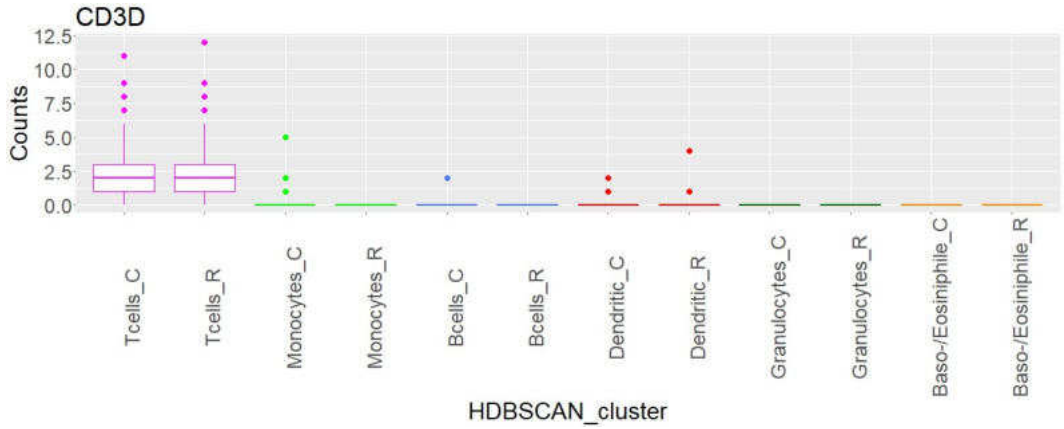


Figure 33. CD3D gene distribution among recognized subpopulations and data samples.

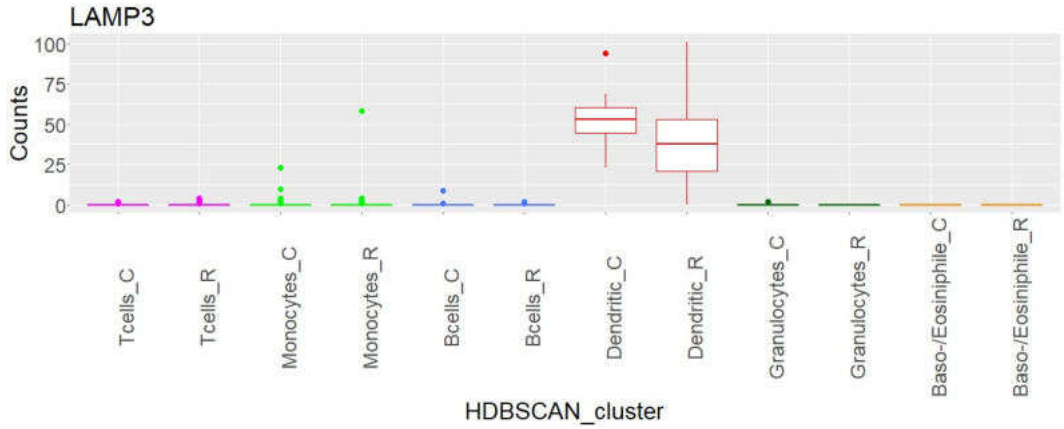


Figure 34. LAMP3 gene distribution among recognized subpopulations and data samples.

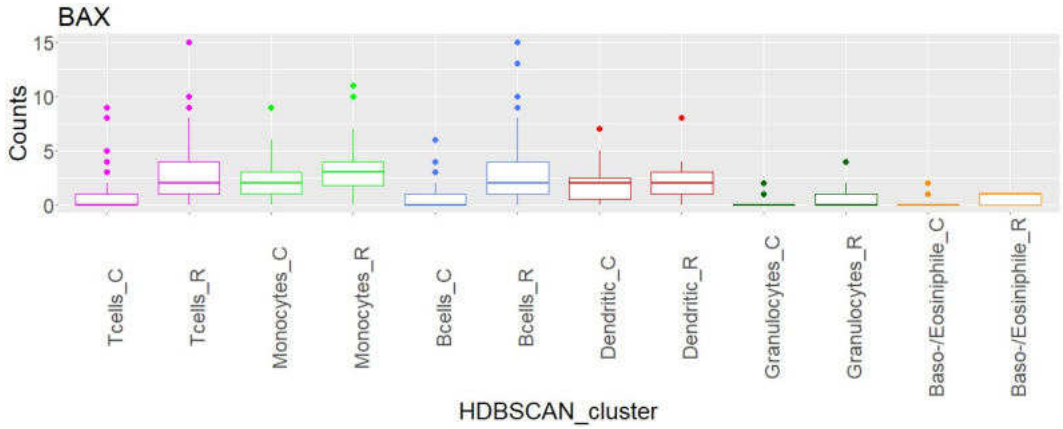


Figure 35. BAX gene distribution among recognized subpopulations and data samples.

Figure 33 and Figure 34 indicate that the CD3D and LAMP3 genes allow the classifier to recognize T-cell and Dendritic subpopulations, respectively. In the case of the CD3D gene, no significant differences can be seen between the count values for the control and irradiated T-cells, further highlighting that this gene was included in the model only due to the intrinsic heterogeneity of the analyzed dataset. On the other hand, in Figure 35, the BAX gene is shown, the presence of which in the model is correctly considered for the differentiation of control and irradiated cells. The distribution of the counts of this gene does not favor any of the detected subpopulations and allows for the detection of differences between the control and irradiated cells. Control and irradiated cell count distribution differences are visible for almost every subpopulation. Therefore, this gene plays a universal role, not considering the dataset's heterogeneity. Considering the above additional analysis steps and the earlier ones concerning the function of the selected set of genes, an undoubted influence of the heterogeneity of the data set on the recognized genetic profile of irradiated cells was found. For this reason, a set of the most numerous T-cell subpopulations was separated to complete the analysis and determine the genetic profile of irradiated cells. Further procedures focused only on this specific part of the data. This made it possible to exclude a solid interfering factor, which is the detected heterogeneity of the dataset caused by the presence of different cell subpopulations. Table 23 shows the observation counts of the T-cell subpopulation for the analyzed data structures.

Table 23. The composition of the test and model structures among T-cell subpopulation.

		Set A	Set B
Test structure	Control	-	429
	Irradiated	-	349
	Total	-	778
Model structure	Control	1051	1540
	Irradiated	801	1330
	Total	1852	2870

The irradiated cells' genetic profile identification procedure remains analogous to the previously described workflow and focuses on set B data at the initial analysis stages. What is more, this analysis is based on the examination of the normalized data. Normalization of observations in a data set without an internal source of variability (caused by other than the expected variability related to the ionizing radiation factor) is essential in the case of the analyzed data. The main factor in favor of its necessity is the occurrence of count values of individual cells in different scales, i.e., different ranges of count

values. Normalization is a process that enables the direct comparison of features over a dataset. The approach to normalizing the data deviates from the standard method of considering the mean and standard deviation but is much less sensitive to outliers [66]. All generated data structures have been subjected to separate normalization procedures. For each data structure, a specific median value was determined for each gene separately (M_g) over all cells from the control sample. The value of Median Absolute Deviation (*MAD*) (22) or Mean Absolute Deviation (*MeanAD*) (24) was also determined depending on the M_g value. Also, all cell count values from the control sample for a particular gene were considered in this case. These two values of normalization parameters, i.e., M_g and *MAD* or *MeanAD*, were used in the normalization process. It was carried out for each control and irradiated cell according to the formulas in equations (21) and (23). Moreover, the numerical values in the formulas used serve as calibration factors to the assumed data distribution, in this case, to the standard normal distribution. It converts the *MAD* to a standard deviation assuming a normal distribution [67]. The value 1.4826 is the inverse of the cumulative distribution function, called the probit function for a normal distribution. More precisely, this value is the reciprocal of the 75% quantile for the standard normal distribution because, for this quantile, 50% of the standard normal cumulative distribution function is covered [68]. When MAD is 0, the scaling factor is 1.2533, which is equivalent to $\sqrt{\frac{\pi}{2}}$ [68].

$$robust\ z_{score}(i,j) = \frac{1}{1.4826} \times \frac{x(i,j) - M_g(j)}{MAD(j)} \quad (21)$$

Where:
i is the row index (cell-specific)
j is column index (gene-specific)
x(i,j) is specified cell count
$M_g(j)$ is the estimated median value for a particular jth gene
MAD(j) is determined by:

$$MAD(j) = M(\left|x(i,j) - M_g(j)\right|) \quad (22)$$

If the MAD(j) value was estimated to be equal to 0, then:

$$robust\ z_{score}(i,j) = \frac{1}{1.2533} \times \frac{x(i,j) - M_g(j)}{MeanAD(j)} \quad (23)$$

Where:
i is the row index (cell-specific)
j is column index (gene-specific)
x(i,j) is specified cell count
$M_g(j)$ is the estimated median value for a particular jth gene
MeanAD(j) is determined by:

$$MeanAD(j) = mean(\left|x(i,j) - M_g(j)\right|) \quad (24)$$

After the normalization process, 50 models were built utilizing the implemented algorithm based on normalized T-cell data. Table 24 summarizes the obtained weighted accuracy values for the validation sets.

Table 24. A set of weighted accuracy values for the validation sets for 50 generated models based on the normalized set B model structure T-cells dataset.

Model ID	Weighted accuracy [%]	Model ID	Weighted accuracy [%]
1	94.37	**26**	96.82
2	94.37	**27**	93.59
3	95.21	**28**	91.94
4	95.16	**29**	100.00
5	97.69	**30**	95.35
6	91.98	**31**	93.59
7	89.61	**32**	95.98
8	94.37	**33**	97.59
9	94.46	**34**	91.98
10	95.98	**35**	95.16
11	96.82	**36**	96.82
12	93.73	**37**	95.16
13	93.55	**38**	98.44
14	92.33	**39**	95.21
15	96.82	**40**	97.59
16	94.37	**41**	97.59
17	98.44	**42**	95.98
18	98.39	**43**	94.65
19	96.08	**44**	93.55
20	97.59	**45**	93.02
21	95.21	**46**	96.08
22	97.69	**47**	96.82
23	95.98	**48**	94.46
24	97.59	**49**	92.84
25	95.98	**50**	95.98

Concerning this data structure, the mean value of the weighted accuracy was 95.40%, and the median value was 95.28%. A minimum weighted accuracy value of 89.61% was achieved, and a maximum value for one of the models of 100.00%. Among the generated models, considered in the context of feature selection as a list of genes, there were 54 unique features. Each gene that occurred at least once in 50 models was assigned a corresponding *GeneRank* value, consistent with (16). A cut-off threshold value was established for the number of significant features, determined in Figure 36. As before, this value was determined based on no changes in subsequent values of the defined metric to the right of the threshold value of the number of features.

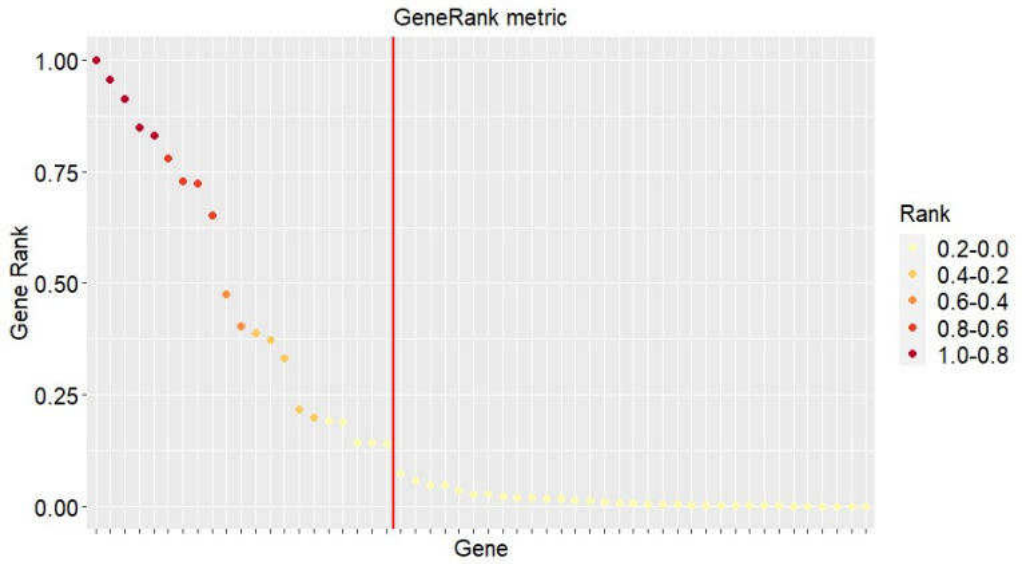


Figure 36. List of features sorted in terms of informativeness with a marked cut-off point for the number of selected features based on the normalized set B model structure T-cells dataset.

The list of selected 21 genes and the corresponding *GeneRank* metric values are presented in Table 25.

Table 25. Complete list of genes selected for constructing the final model based on the normalized set B model structure T-cells dataset.

Gene	GeneRank	Gene	GeneRank
RPS19P1	1.0000	CD74	0.3878
RPL23AP42	0.9564	TNFSF8	0.3742
BAX	0.9127	THBS1	0.3316
DDB2	0.8491	STAT1	0.2183
HLA-A	0.8307	TRIB2	0.2001
RPS27L	0.7792	LCK	0.1922
PHPT1	0.7277	AEN	0.1904
MYC	0.7224	CDKN1A	0.1448
CCNG1	0.6509	STAT51	0.1429
FYB	0.4759	CHI3L1	0.1407
CD52	0.4033		

The estimated final model parameter values for 21 chosen features are presented in Table 26. The model parameters were calculated based on the set B model structure, consisting of 2870 normalized T-cells.

Table 26. Estimated parameter values for the final model based on the normalized set B model structure T-cells dataset.

Intercept	RPS19P1	RPL23AP42	BAX	DDB2	HLA-A	RPS27L	PHPT1
-4.47	1.76	-1.44	0.60	0.22	-0.75	0.40	0.32
MYC	**CCNG1**	**FYB**	**CD52**	**CD74**	**TNFSF8**	**THBS1**	**STAT1**
-0.54	0.32	-0.39	-0.39	-0.22	0.10	-0.30	-0.31
TRIB2	**LCK**	**AEN**	**CDKN1A**	**STAT5A**	**CHI3L1**		
-0.32	-0.33	0.06	0.05	-0.29	0.11		

The final stage of building the complete model was determining the new probability threshold value for irradiated cells classification utilizing the previously described Youden index (17). Results are presented in a graphical form, with the 0.5124 new classification probability threshold value marked in Figure 37. As in the previously described approach, this value was determined based on a set B test structure.

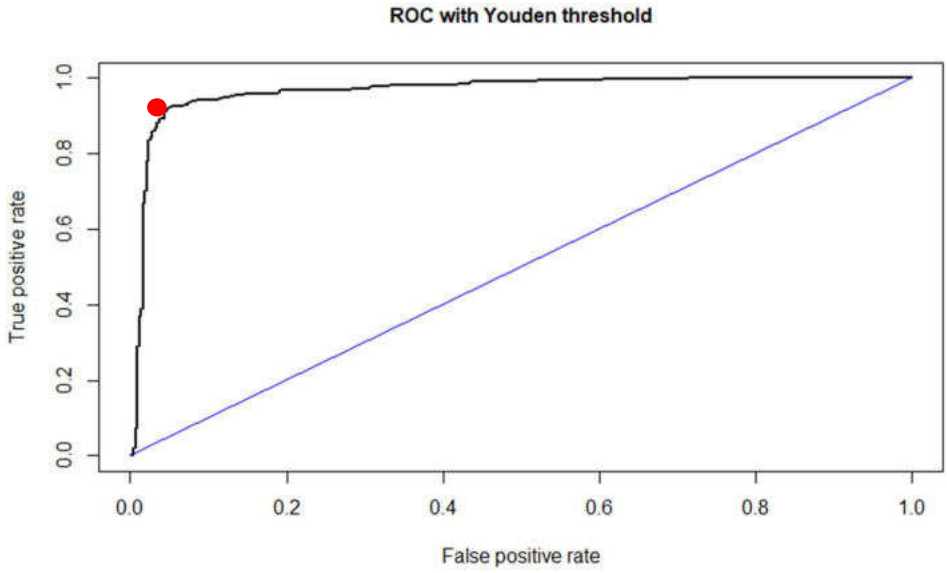


Figure 37. ROC with marked Youden threshold for the final model.

The effectiveness of utilizing the new probability threshold value for control and irradiated cells classification purposes was compared against a predetermined value of 0.50 classification probability value. These results are included in Table 27.

Table 27. Comparison of the classification quality metrics for the two irradiated cells classification probability threshold values.

	Classification threshold	
Quality metric name	**Fixed 0.5000**	**Youden 0.5124**
TP	323	323
TN	405	407
FP	24	22
FN	26	26
Precision	0.9308	0.9362
Sensitivity	0.9255	0.9255
Specificity	0.9440	0.9487
Weighted accuracy	0.9357	0.9383
$F1_{score}$	0.9282	0.9308
Number of cells	778	778
Number of correctly classified cells	728	730
Number of incorrectly classified cells	50	48
Incorrectly classified cells [%]	6.43	6.17

As a result of applying a slightly higher classification probability threshold value, two additional control cells were correctly classified. In this case, a high value of the weighted classification accuracy of 93.83% was achieved. This means that only 48 out of 778 cells were not correctly classified. Also, in this case, applying a new probability threshold value is reflected in the higher $F1_{score}$ metric value. In addition, independent testing was carried out based on the set A model structure T-cells data. Before testing, this data structure was also normalized following the procedures carried out for the data structures of set B. Results for independent testing of the final model are shown in Table 28.

Table 28. Classification quality metric values based on the independent test set.

Quality metric name	**Quality metric value**
TP	751
TN	989
FP	62
FN	50
Precision	0.9237
Sensitivity	0.9376
Specificity	0.9410
Weighted accuracy	0.9395
$F1_{score}$	0.9306
Number of cells	1852
Number of correctly classified cells	1740
Number of incorrectly classified cells	112
Incorrectly classified cells [%]	6.05

The presented classification quality measures are very high considering, first of all, the value of the weighted classification quality of almost 94% and the high value of the $F1_{score}$ measure of 0.93. Such testing on an independent data set allowed for confirming the high quality and specificity of the detected genetic profile of irradiated cells in the context of the problem of recognizing control and irradiated cells.

6.4 Irradiated cells' genetic profile recognition summary

The proposed and applied workflow for the problem of genetic profiles of cells irradiated in an ex vivo environment based on data from single-cell sequencing experiments recognition allowed the detection of the characteristic structure of cells irradiated with a dose of 1 Gy. An exceptionally high $F1_{score}$ value estimated at 0.9308 indicates satisfactory classification results concerning control and irradiated cells. Another essential aspect that allows determining the usefulness of the recognized model of irradiated cell structure is the reduction of dimensionality, in other words, the number of features contained in this model. This aspect has undoubtedly been successfully implemented, guaranteeing a significant decrease in the number of informative features, selecting only 21 genes out of 406 that were considered at this stage of the analysis. These selected genes form the structure that enables the correct classification of cells and gives an answer to the stated goal of this work, i.e., recognition of the genetic profile of cells irradiated in an ex vivo environment. Moreover, in the case of the detected genetic profile, an analogous list of genes and their functions was performed, as in the *White blood cell subpopulations recognition* subsection. The results are presented in Table 29.

Table 29. Genes of the recognized profile of irradiated cells with their corresponding functions.

Gene name	Function/process
RPS19P1	-
RPL23AP42	-
BAX	Apoptosis regulator
DDB2	-
HLA-A	Housekeeping
RPS27L	-
PHPT1	-
MYC	Proliferation marker
CCNG1	-
FYB	Miscellaneous
CD52	CD marker
CD74	Cell type marker
TNFSF8	Cytokine
THBS1	Cell adhesion
STAT1	Transcription factor
TRIB2	Kinase
LCK	Marker gene
AEN	-
CDKN1A	-
STAT5A	Transcription factor
CHI3L1	Enzyme

One crucial question should be answered: how can we decide that the recognized structure contains the necessary radiation response genes in the posed problem? In this case, the answer can be directly provided by a detailed study concerning the analysis of both the functions of individual genes

(functional analysis) and determining which of the selected genes have already been described as radiation response genes in previously published works (signature analysis). The purpose of the functional analysis on the selected set of genes is primarily to determine whether these genes, due to their functions, can be directly or indirectly related to the phenomenon of irradiation. The functional analysis thus provides biological insight into the hypothesis that the genetic profiles of irradiated cells are altered compared to control cells. Therefore, we expect enrichment in the pathways responsible for the cellular response to ionizing radiation or the response to damage to the genetic material (translation, mRNA regulation). The *enrichGO()* [68] function from the *clusterProfiler* package in the R environment was used to perform the functional analysis. The results of the functional analysis are presented in Figure 38. Only biological processes with a determined adjusted p.value < 0.0025 are shown for clarity and readability.

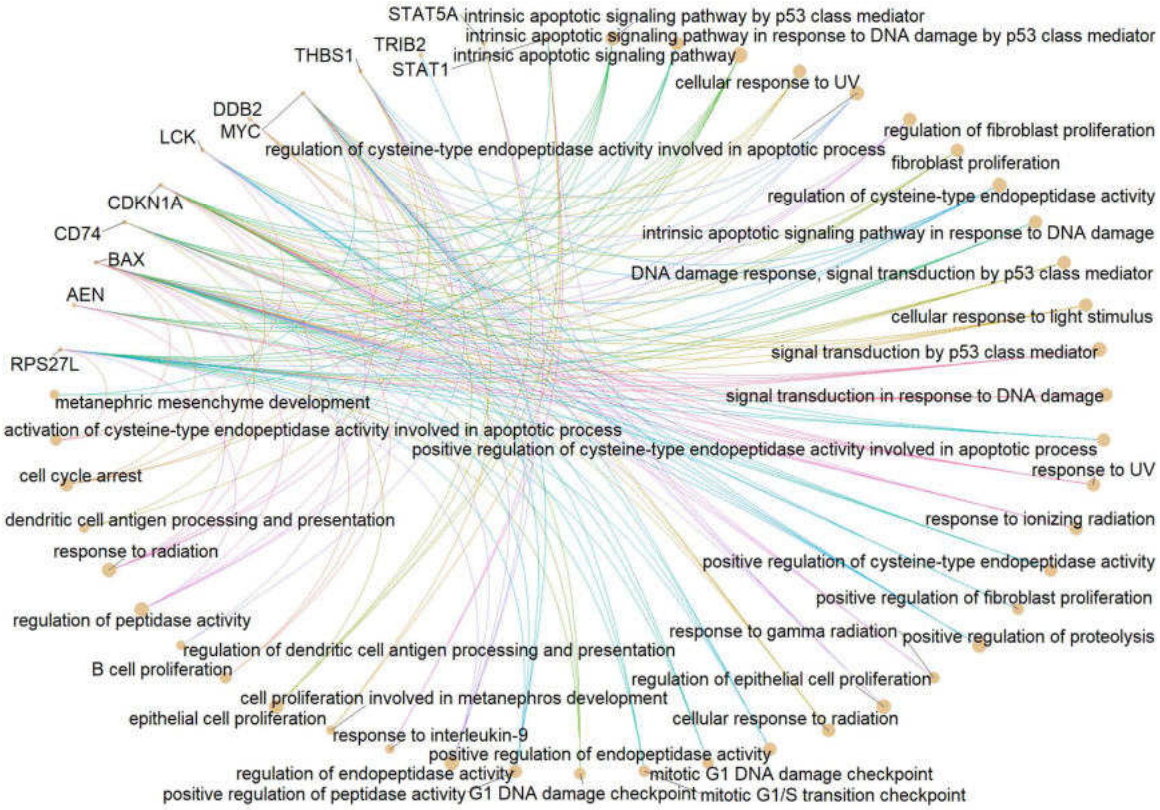


Figure 38. Functional analysis results for 21 selected features with marked significant BPs.

Additionally, in Table 30, the radiation-related BPs are indicated, and the adjusted significance threshold considered equals 0.01.

Table 30. The number of significant genes (with adjusted p.value < 0.01) connected with radiation-based BP paths.

BP connected with irradiation	Number of genes (p.value)
intrinsic apoptotic signaling pathway	6 (0.000086)
intrinsic apoptotic signaling pathway by p53 class mediator	5 (0.000013)
response to radiation	5 (0.001551)
lymphocyte proliferation	4 (0.003594)
positive regulation of cysteine-type endopeptidase activity involved in the apoptotic process	4 (0.000548)
G1/S transition of mitotic cell cycle	4 (0.002808)
cell cycle G1/S phase transition	4 (0.003150)
mononuclear cell proliferation	4 (0.003594)
cellular response to environmental stimulus	4 (0.003731)
leukocyte proliferation	4 (0.004261)
cellular response to UV	4 (0.000224)
cellular response to abiotic stimulus	4 (0.003731)
response to UV	4 (0.000588)
cellular response to radiation	4 (0.001026)
regulation of apoptotic signaling pathway	4 (0.008370)
signal transduction in response to DNA damage	4 (0.000526)
response to ionizing radiation	4 (0.000647)
intrinsic apoptotic signaling pathway in response to DNA damage	4 (0.000305)
DNA damage response, signal transduction by p53 class mediator	4 (0.000305)
cellular response to light stimulus	4 (0.000391)
cellular response to tumor necrosis factor	4 (0.004490)
activation of cysteine-type endopeptidase activity involved in the apoptotic process	3 (0.001832)
mitotic G1 DNA damage checkpoint	3 (0.001259)
response to gamma radiation	3 (0.001026)
mitotic DNA damage checkpoint	3 (0.002808)
mitotic DNA integrity checkpoint	3 (0.002972)
T cell receptor signaling pathway	2 (0.009061)

Detailed data and information on all detected statistically significant (adjusted p.value < 0.01) biological processes are presented in tables in the *Additional materials* section in the subsection *Ex vivo data functional analysis.*

The signature analysis was based on a literature review of sources that would indicate an earlier direct or indirect connection with the phenomenon of radiation. Table 31 presents the genes identified in the performed analysis workflow as important in the problem of distinguishing control and irradiated cells, with marked (green) genes described in literature sources as radiation response genes.

Table 31. Signature analysis results for recognized irradiated cells' genetic profile.

Gene number	Gene name
1	RPS19P1 [83]
2	RPL23AP42 [83]
3	BAX [82]
4	DDB2 [85]
5	HLA-A
6	RPS27L [84]
7	PHPT1 [87]
8	MYC [91]
9	CCNG1 [97]
10	FYB [93]
11	CD52
12	CD74 [102]
13	TNFSF8 [96]
14	THBS1 [103]
15	STAT1 [104]
16	TRIB2
17	LCK [99]
18	AEN [84]
19	CDKN1A [96]
20	STAT5A [98]
21	CHI3L [105]

A thorough analysis of the genes' biological functions revealed a strong relationship with the phenomenon of response to ionizing radiation. The pathways of biological processes described included, among other things, the cellular response to UV and the response to ionizing radiation. There was found that 5 out of 21 detected genes were directly related to the response to radiation factor. In addition, recognized genes essential in response to radiation showed several processes directly associated with this phenomenon and the enormity of biological pathways related to damage to the genetic material or general mobilization due to an external factor. These pathways, taking into account the abundance of the detected genes, mainly include processes related to apoptotic communication, cellular response to an environmental stimulus, DNA damage-checking pathways, and lymphocyte proliferation. Additionally, the signature analysis shows a very high coverage with the current literature reports regarding the radiation response genes. Only 3 of 21 genes were not reflected in the available scientific articles. Determining the functions performed by each of these three additional genes utilizing the *GeneCards* [69] platform enabled the determination of these other three genes in the context of the radiation response. HLA molecules play a vital role in the immune system. They are responsible for the presentation of peptides, which, in turn, can be recognized by cytotoxic T lymphocytes. CD52 is responsible, among other things, for regulating the concentration of calcium ions in the intracellular environment. In turn, TRIB2 is associated with activating and modulating signaling pathways in physiological and pathological processes. According to specific genes' functions, their expression can be modified due to an external factor acting on them.

Only one question remains: is the model built with 21 features much better than the model built based on only one most significant feature? Is making a complex model in the context of the achieved classification quality metrics profitable? To respond, an uncomplicated analysis was carried out based on comparing the created Multiple Input Single Output (MISO) model and 21 Single Input Single Output (SISO) models. This analysis was carried out in several successive steps. The first one was estimating the values of the parameters of the one-factor models based on the model structure of the B dataset. Then, a new classification probability value was recalculated for each model using the Youden index. In the last step, the obtained classification results were tested utilizing the set B test structure containing 778 cells. Table 32 shows the results of this consideration.

Table 32. MISO vs. SISO analysis preformed for main model from ex vivo data analysis.

Model	Youden threshold	Precision	Sensitivity	Specificity	Weighted accuracy	F1 score	Parameters
RPS19P1	0.4672	0.8017	0.8271	0.8353	0.8316	0.8142	-0.6949; 1.5454
RPL23AP42	0.4506	0.4425	0.6657	0.3248	0.4769	0.5316	-0.1823; -0.0558
BAX	0.8012	0.8897	0.6974	0.9304	0.8265	0.7819	-1.8462; 0.8686
DDB2	0.7296	0.8170	0.5533	0.9002	0.7455	0.6598	-0.7938; 0.2607
HLA-A	0.4695	0.4973	0.5360	0.5638	0.5514	0.5159	-0.1939; -0.3198
RPS27L	0.7410	0.8058	0.5620	0.8910	0.7442	0.6622	-1.5205; 0.8511
PHPT1	0.7452	0.7790	0.6196	0.8585	0.7519	0.6902	-0.9105; 0.4581
MYC	0.4950	0.4724	0.8127	0.2691	0.5116	0.5975	-0.0200; -0.3291
CCNG1	0.6523	0.8507	0.5418	0.9234	0.7532	0.6620	-1.1435; 0.6114
FYB	0.4582	0.4435	0.5879	0.4060	0.4871	0.5056	-0.1675; -0.0634
CD52	0.4480	0.4700	0.6772	0.3852	0.5154	0.5549	-0.2086; -0.2668
CD74	0.4868	0.5448	0.6311	0.5754	0.6003	0.5848	-0.2185; -0.2455
TNFSF8	0.6245	0.7384	0.3660	0.8956	0.6594	0.4894	-0.5601; 0.1736
THBS1	0.8278	0.0000	0.0000	1.0000	0.5540	0.0000	-0.1692; 0.0155
STAT1	0.4804	0.4582	0.6484	0.3828	0.5013	0.5370	-0.0783; -0.1899
TRIB2	0.4182	0.4543	0.9308	0.0998	0.4704	0.6106	-0.1139; -0.1211
LCK	0.4697	0.4737	0.5187	0.5360	0.5283	0.4952	-0.1875; -0.0981
AEN	0.7267	0.8358	0.3228	0.9490	0.6697	0.4657	-0.5107; 0.0992
CDKN1A	0.0010	0.4453	0.9971	0.0000	0.4447	0.6157	-0.1918; -0.0664
STAT5A	0.4702	0.4503	0.6398	0.3712	0.4910	0.5286	-0.1193; -0.1195
CHI3L1	0.9918	1.0000	0.0231	1.0000	0.5643	0.0452	-0.1790; 0.0536
Full model	0.5124	0.9362	0.9255	0.9487	0.9383	0.9308	-

The constructed model, consisting of 21 features, was marked as a *Full model*, while the remaining 21 single-feature models are marked with the subsequent gene names. This table compares all the most crucial classification quality metrics utilized in this dissertation. Thus, it is possible to answer the question posed. The MISO model is much more advantageous than all other SISO models. By comparing the $F1_{score}$ metric, no other than the MISO one exceeds the value of 0.90 for this metric. For this model, the $F1_{score}$ metric value is above 0.93. Due to the highest $F1_{score}$, it is also possible to infer the best precision-sensitivity ratio for data classification. Moreover, the *Full model* achieved a weighted accuracy value of 93.83%, while the best SISO model (RPS19P1) achieved a much lower value of 83.16%. This unambiguously determines the superiority of the multi-factor model over the one-factor models, also considering the number of included features.

7 Ex vivo irradiated cells' genetic profile recognition based on the neural networks

The second machine learning approach was also aimed at determining the genetic profile of irradiated cells and the classification of irradiated and control cells based on the built model. It was necessary to apply feature selection methods beforehand. In this case, feature selection based on neural network theory was applied. However, the feature selection was used primarily with filter methods and not, as in the case of logistic regression, first wrappers and then filters. The approach utilized in building a model based on LR is much more time-consuming precisely because of the use of computationally expensive wrappers on a complete set of features in the first line. In the case of the neural network, filters were first applied on a complete set of genes and then wrappers to a minimal set of selected features. This part of the doctoral dissertation is based on a previously selected and normalized subset of T-cells subpopulation data. The division into a test set, which does not participate in the processes of building the classifier structure, and a model set remains valid, as previously described. The numerical content of control and irradiated cells is included in Table 23. The conducted analysis consisted of several main aspects, such as the selection of the structure of the classifier based on neural network methods, the features selection using the filter method, building a model based on the chosen set of informative genes, and testing the model both on the test structure of data set B and an independent model structure of data set A.

The appropriate neural network model structure research required the analysis of the impact of several characteristics as the number of neurons in the subsequent layers, the dropout level, the number of learning epochs, and the number of layers built into the neural network. Each neural network must have one input layer that accepts features and one output layer that transmits the results. Therefore, the number of neurons in the input layer equals the number of features to be analyzed. In contrast, the output layer in the presented workflow consists of only one element, providing information about the observations' probability of belonging to the irradiated sample. In hidden layers, the number of neurons is determined based on the formula (25):

$$N_{neurons} = \frac{N_{inputs} + N_{outputs}}{2} \quad (25)$$

The dropout level in hidden layers was set a priori to 10% of the number of neurons in a given layer. Dropout is essential when learning complex neural networks on large and complicated data sets. By rejecting a specific level or number of neurons from the network learning process in a given layer, dropout prevents too good matching to this data set, significantly reducing or even preventing the network from overfitting [70] [71]. This is a very undesirable phenomenon because most models are built for later use on other data sets. In the case of an overtrained model, it does not meet the assumed goals of generalization of the problem and is only fitted to the training data. The next, worth emphasizing term is the training epoch. It is described as one complete training cycle based on the entire training set. Therefore, the number of training epochs defines how many times the training algorithm will go through the training set [72]. In the case of the presented workflow, the number of learning epochs is only a conventional element to shorten the calculation time. It is treated as the maximum number of learning epochs rather than choosing the optimal number after which the model parameters will be delivered. The model parameters are selected based on the highest value of the $F1_{score}$ metric achieved based on the validation set among all training epochs. The complete set of features concerning the model structure of data set B was analyzed to determine the maximum number of

training epochs. The analyzed model structure consisted of 3 layers, including one hidden layer. Results are presented in Figure 39.

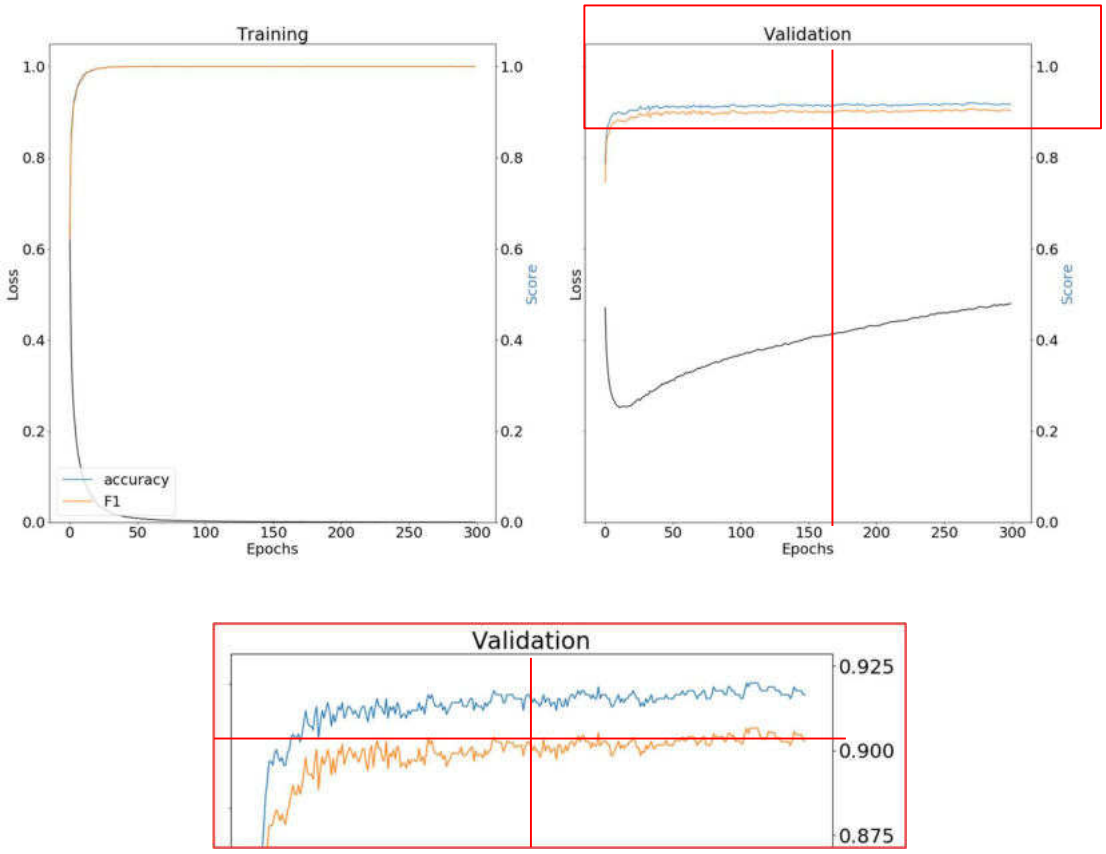


Figure 39. Selection of the number of training epochs for the model of neural networks with three layers.

After passing 150 training epochs, there is a visible and long-term stabilization of the $F1_{score}$ metric value for the validation set. Based on the performed analysis, the maximum number of training epochs was selected and set at 150. The last step of building the general structure of the neural network model is the number of network layers selection. Three variants containing the network's 3, 5, and 10 layers were analyzed. The analysis was performed over 150 training epochs based on the model structure of dataset B. The results are shown for the training and validation data structure in Figure 39, Figure 40, and Figure 41. Moreover, the number of layers equal to 10 means one input layer, eight hidden layers, and one output layer. This scheme applies to the other analyzed variants.

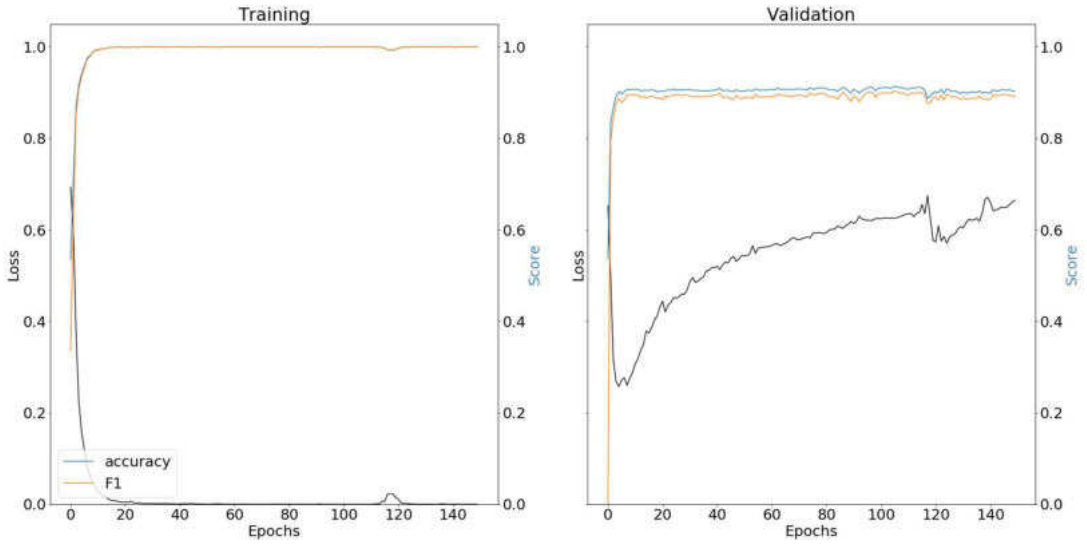


Figure 40. Selection of the number of neural network layers for the model with five layers.

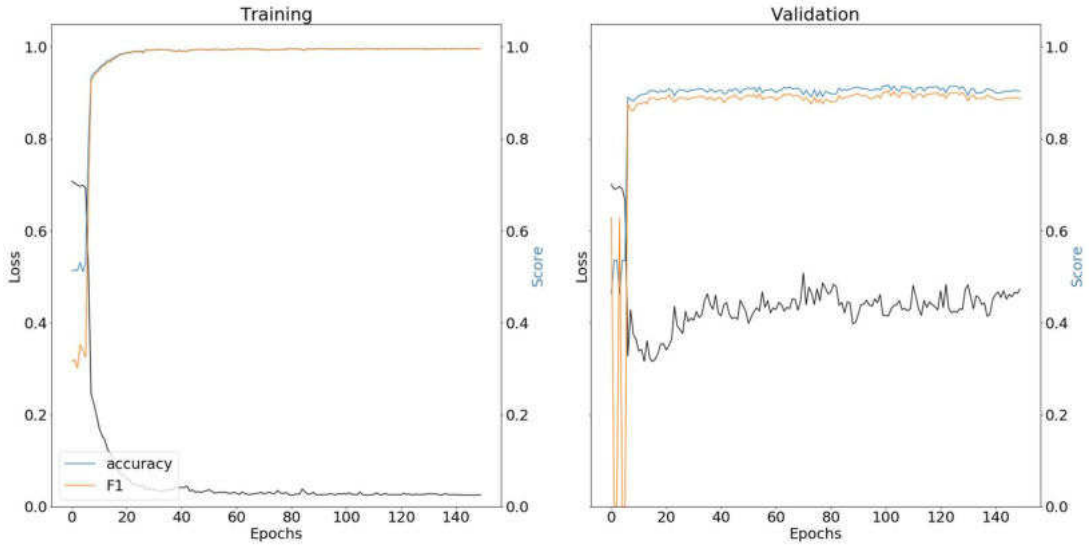


Figure 41. Selection of the number of neural network layers for the model with ten layers.

The decision about the number of layers for the given problem was also based on the value of the $F1_{score}$ metric for the validation set. The differences between the successive numbers of layers are minor, taking into account the value of this metric, while two aspects support the choice of three network layers. The first is the highest value of the $F1_{score}$ metric achieved for this approach, and the second is the lowest degree of complexity of the model structure. For this reason, three layers were selected to create a neural network model. A summary of the final neural network model details is presented in Table 33.

Table 33. Details of the neural network model structure.

Analyzed feature	Applied
Number of neurons	Equation (25)
Dropout level	10% neurons
Number of learning epochs	150
Number of layers	3 (1+1+1)

The structure of the neural network model is undoubtedly an essential part of the observations' classification problem. Nevertheless, as with a logistic regression model, or any other model designed to classify observations correctly, feature selection is one of the most critical aspects. In the earlier chapters of this dissertation, attention was drawn to the benefits and risks of feature selection and the need for its usage in the case of high-dimensional data. This line of reasoning was also adopted in this case. In the feature selection process, the retention of information about possible correlations between features was prioritized. For this reason, a model of neural networks was built on the model structure of data set B for all 406 available features. For this purpose, the function *Sequential()* from *TensorFlow.Keras* library [73] was used. To maintain the possibility of choosing the best set of features, the *ModelCheckpoint()* function from the same library in the Python environment was utilized. The introduced functionality allows monitoring the changing parameters of the model on an ongoing basis and saving only those that meet the given condition. In the analysis, this condition was to obtain the highest value of the $F1_{score}$ metric based on the training set. The model was trained using the *fit()* function for a given training set with observations sample of origin information, the number of epochs set to a predetermined value of 150, and a validation set with annotations of the observations sample of origin information. The utilized function returns the values of the specified metric for both input datasets. Moreover, the validation set is not directly involved in training the model. Still, it is only used to calculate the values of the returned metrics to give a broader insight into the models' performance. Based on the validation set, these values are fundamental when determining the degree of generalization of the model being built or for observing and reacting appropriately in the event of model overfitting. This phenomenon is very well visible in the previously presented Figure 39, Figure 40, and Figure 41, where in most cases, an increase in the value of the loss function can be observed (black line). An increase in the loss function with no decrease in the classification quality could indicate the classification uncertainty level increase.

After fitting the model to the given training set and returning the values of the neural network model parameters, it was possible to proceed directly to the feature selection part of the workflow. For this purpose, the popular and widely used *shap* [74] library was utilized to explain the 'black box' problem in neural networks. Moreover, this method is recommended and used more often for feature selection [75] [76]. This tool's primary goal is to determine each feature's influence on the classification of a single observation. The methodology included in this tool enables computing Shapley values based on coalition game theory to distribute the features' contributions fairly. Therefore, this tool plays a vital role in using neural networks, which until now were considered mentioned 'black boxes' because their actions and decisions could not be fully explained. The interpretation itself is critical in the growing social and scientific expectations, where an increasingly conscious society expects a full explanation of the functioning and dependencies of a given problem. The *Shapley Additive exPlanations* (SHAP) tool's fundamental task is to interpret the machine learning model in the context of model learning and prediction results [76]. The concept of Shapley values is not new, as it was already described in 1953 by Lloyd S. Shapley [77] [78]. However, it is particularly appreciated in this case, and this approach enables interpreting the utilized neural network model. The very idea of the significance of SHAP features is based on the assessment that the higher the absolute value assigned to a specific feature, the

more important this feature is. To carry out a global interpretation of the features' influence, there have to be considered the absolute value of the feature over the entire data [76] (26):

$$S = \frac{1}{N}\sum_{i=1}^{N} |ShapValue_i| \qquad (26)$$

Where:
N is the number of observations in the dataset

As a result of applying the described procedures, a measure of the significance of individual features was obtained called *ShapScore*. The results are presented in Figure 42 for a set of 406 analyzed input features.

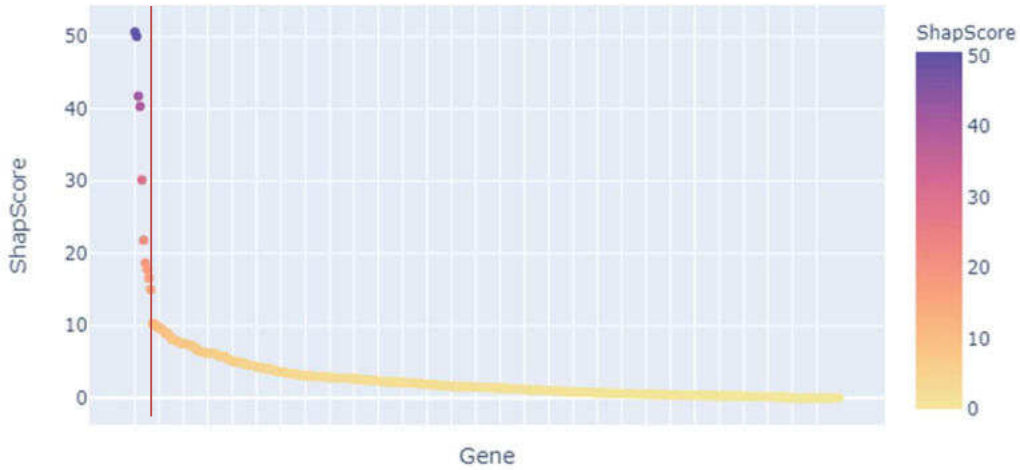


Figure 42. List of importance-sorted features with the marked cut-off point for the number of informative genes based on *ShapScore*.

The previously described methodology was used to determine a threshold value for the number of features based on *ShapScore*. Table 34 represents a list of 10 features with the assigned *ShapScore* values.

Table 34. *ShapScore* values for chosen informative genes.

Gene	ShapScore value
RPS19P1	50.62
BAX	50.02
DDB2	41.75
PHPT1	40.34
RPS27L	30.15
CCNG1	21.83
AEN	18.69
PCNA	17.78
TNFRSF10B	16.57
TNFSF8	14.98

Based on the ten chosen genes, the final model was built with the estimated parameter values. Figure 43 shows the course of the training process with internal validation based on the model structure of data set B.

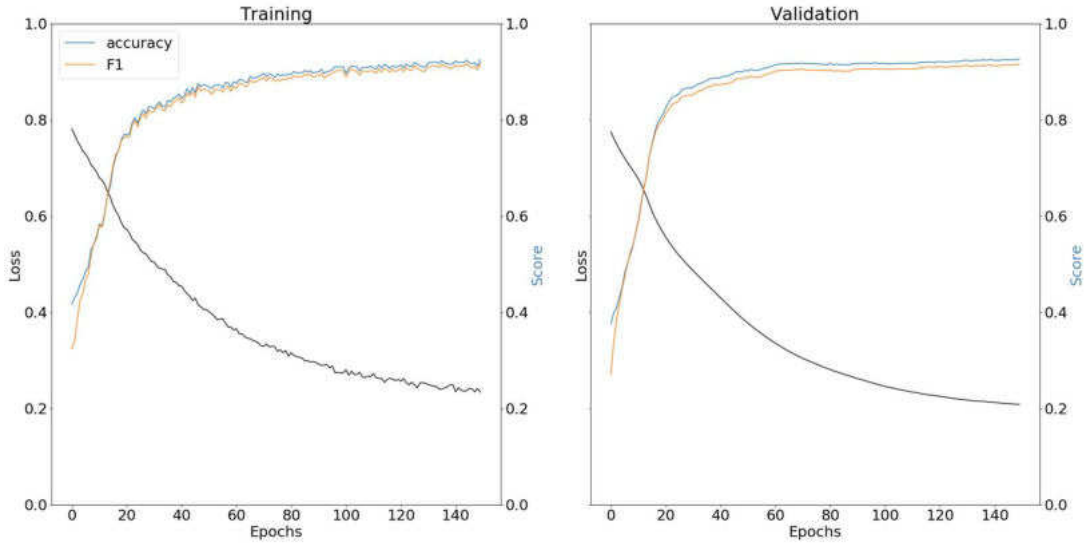


Figure 43. The course of learning and validation of the neural networks-based model.

What is very important, as a result of the features selection and the use of selected genes in the final formula of the model based on neural networks, it became possible to eliminate the phenomenon of the increase in the loss function, which was directly related to the decrease in classification certainty and the phenomenon of model overfitting. The loss function takes lower and lower values for both the training and validation sets, while the classification quality and $F1_{score}$ metric increase over 150 training epochs. The created neural network-based model was subjected to a test procedure based on the test structure of data set B. The exact values of the classification quality metrics are presented in Table 35.

Table 35. Classification quality metric values based on the set B test structure.

Quality metric name	Quality metric value
TP	308
TN	402
FP	27
FN	41
Precision	0.9194
Sensitivity	0.8825
Specificity	0.9371
Weighted accuracy	0.9098
$F1_{score}$	0.9016
Number of cells	778
Number of correctly classified cells	710
Number of incorrectly classified cells	68
Incorrectly classified cells [%]	8.74

The classification quality metrics achieved by the classifier are satisfactory, as evidenced by high precision and specificity values above 0.90 and a sensitivity value above 0.88. The classifier performs very well in recognizing control cells and slightly worse in identifying irradiated cells, as evidenced by the higher level of FN cases, which negatively affects the sensitivity measure. The classification's weighted accuracy was determined at almost 91%, with a high value of the $F1_{score}$ measure above 0.90. The overall percentage of misclassified cells was over 8.50%.

The last step was to conduct independent testing based on the model structure of dataset A. The classification quality metrics based on the independent testing are shown in Table 36.

Table 36. Classification quality metric values for independent testing based on the set A model structure.

Quality metric name	Quality metric value
TP	738
TN	976
FP	75
FN	63
Precision	0.9077
Sensitivity	0.9213
Specificity	0.9286
Weighted accuracy	0.9250
$F1_{score}$	0.9087
Number of cells	1852
Number of correctly classified cells	1714
Number of incorrectly classified cells	138
Incorrectly classified cells [%]	7.45

Independent testing allowed for very high measures of the three described classification quality metrics, precision, sensitivity, and specificity, each above the value of 0.92. A weighted classification accuracy of over 92% was also achieved. The percentage share of incorrectly classified cells was below 7.50%, proving satisfactory classification results based on the created neural network model. Also, classification quality metrics are better for the independent test set than the internal one.

8 Logistic Regression and Neural Networks - results comparison

Applied machine learning workflows enabled the detection of the appropriate genetic structures of irradiated cells and the creation of classifiers to separate control and irradiated cells by assigning suitable classes. In the case of using logistic regression methods to determine the genetic profile of irradiated cells, a more complex structure was obtained, consisting of 21 selected genes. In the case of using neural network methods, the recognized genetic structure was twice as small, including only ten genes. Genetic profiles in both instances of the machine learning methods were based on analyzing an identified subset of T-cells over the white blood cells. All machine learning procedures were performed based on the model structure of data set B. The testing procedures in both cases were based on the test structure of data set B, which was removed from the analysis at an early stage. Thus, this structure did not affect the selection of individual genes and the processes of creating the final model for cell classification. Therefore, testing was carried out on the same data set concerning the cellular composition for both methods, enabling their direct comparison.

Selected genes for both utilized approaches are shown in Table 37. Moreover, relevant literature references are provided for radiation response genes.

Table 37. Irradiated cells' genetic signature recognized based on logistic regression and neural networks approaches.

Logistic Regression-based		Neural Networks-based	
Gene	**Radiation response**	**Gene**	**Radiation response**
RPS19P1	Radiation response [83]	RPS19P1	Radiation response [83]
RPL23AP42	Radiation response [83]	BAX	Radiation response [82]
BAX	Radiation response [82]	DDB2	Radiation response [85]
DDB2	Radiation response [85]	PHPT1	Radiation response [87]
HLA-A	-	RPS27L	Radiation response [84]
RPS27L	Radiation response [84]	CCNG1	Radiation response [97]
PHPT1	Radiation response [87]	AEN	Radiation response [84]
MYC	Radiation response [91]	PCNA	Radiation response [84]
CCNG1	Radiation response [97]	TNFRSF10B	Radiation response [86]
FYB	Radiation response [93]	TNFSF8	Radiation response [96]
CD52	-		
CD74	Radiation response [102]		
TNFSF8	Radiation response [96]		
THBS1	Radiation response [103]		
STAT1	Radiation response [104]		
TRIB2	-		
LCK	Radiation response [99]		
AEN	Radiation response [84]		
CDKN1A	Radiation response [96]		
STAT5A	Radiation response [98]		
CHI3L	Radiation response [105]		

Based on current literature sources, as many as 18 out of 21 genes detected using logistic regression methods are described as radiation response genes. The functions of the other genes have been determined based on the *GeneCards* platform [69]. They include the HLA-A gene that plays a central

role in the immune system (the process of recognition of peptides by cytotoxic T cells), the CD52 gene associated with the regulation of cytosolic calcium ion concentration, and the TRIB2 gene responsible for processes of signaling pathways in physiological and pathological processes. In turn, in the case of genes detected using neural networks, all of them were described as radiation response genes. Moreover, as many as eight detected genes are common to both workflows, as shown in Figure 44.

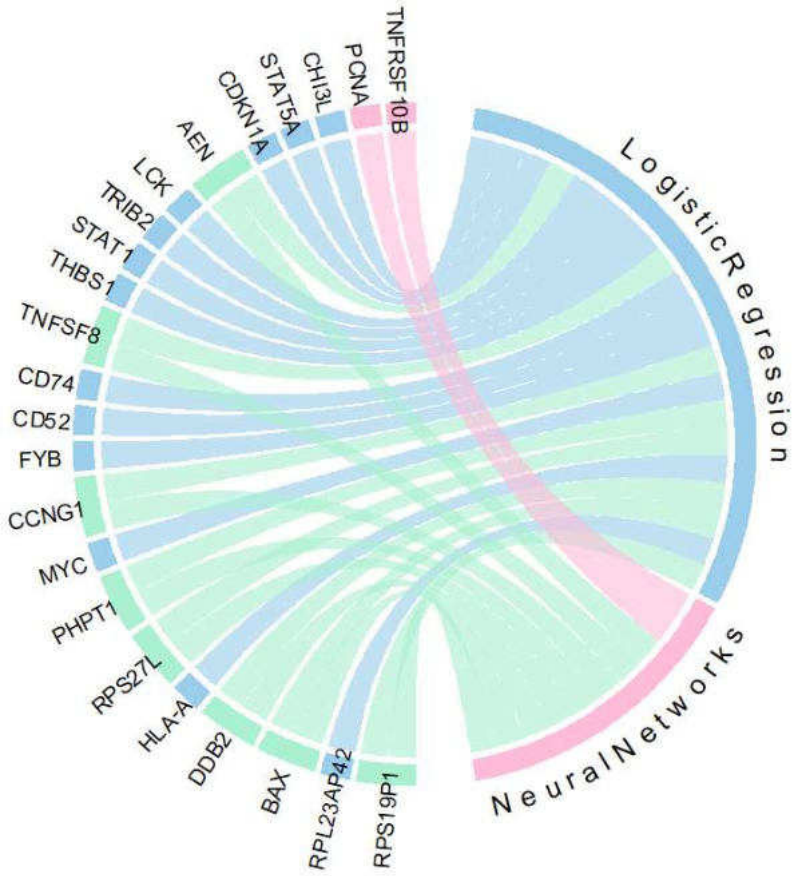


Figure 44. LR and NN-based irradiated cells genetic profiles comparison.

Genes common to both workflows are marked in green, genes characteristic of the approach based on logistic regression in blue, and genes based on neural networks in red.

An additional analysis was also conducted to determine whether one of the two detected genetic profiles of irradiated cells is more universal and whether they can be transferred to another classification method without classification quality loss. Three approaches to feature selection were compared. The first was the lack of feature selection, i.e., building a model on all 406 genes available after quality control for the model structure of data set B. In the second and third variants, genetic structures were used as a result of feature selection using logistic regression and neural network methods. These structures contained 21 and 10 selected genes, respectively. The built models were then tested using the approach based on logistic regression and neural networks, i.e., each model was tested twice, each time based on a different machine learning methodology. In Table 38, there are described three model selection approaches compared between both machine learning methods. The values of the classification quality metrics were determined based on the test structure of dataset B.

Table 38. Comparison of the classification quality for three model selection approaches for both machine learning methods applied.

Model selection approach	Length of signature	Logistic Regression		Neural Networks	
		w. accuracy	$F1_{score}$	w. accuracy	$F1_{score}$
No selection	406	0.8856	0.8712	0.8638	0.8453
Forward Selection for LR	21	0.9383	0.9308	0.9369	0.9305
Shap Scoring (conventional) for NN	10	0.9165	0.9057	0.9098	0.9016

The lack of feature selection for both analyzed machine learning methods resulted in the worst results considering the presented classification quality measures. This model was very complicated and contained as many as 406 features, which required determining the values of model parameters and proceeding with observations on the entire feature space. However, when considering models built based on feature selection for both machine learning approaches, the best results were achieved for the model built based on feature selection using logistic regression methods (*Forward Selection for LR*) and also processed using logistic regression methods. The weighted classification accuracy was almost 94%, with the value of the $F1_{score}$ measure above 0.93. However, for the model built in this way and the classification of cells using neural networks, comparably satisfactory results were achieved, considering the value of the $F1_{score}$ metric above 0.93 and the weighted classification accuracy above 93.5%. However, the model built based on neural network methods achieved significantly worse results than the previously described model based on logistic regression. Both for classification using logistic regression methods and neural networks, the $F1_{score}$ metric value was above 0.90, and the weighted classification quality was about 91%.

9 Conclusions

The objectives of this doctoral dissertation have been successfully achieved, both in terms of detecting white blood cell subpopulations, recognizing ex vivo irradiated cells' genetic profiles and the two different machine learning methods for feature selection application comparison. A specific and complex workflow of single-cell data analysis was developed, including new techniques and publicly available tools and algorithms. An additional advantage of the work is the implementation of an algorithm for the features selection based on the logistic regression methods, which significantly improves, and, what is more, enables the proposed methodology of analysis to be carried out on high-dimensional scRNA-seq data.

9.1 White blood cell subpopulations recognition

The ex vivo data, consisting of two datasets from technical repetitions of the experiment, showed a high heterogeneity detected using the unsupervised UMAP learning methods. Performed procedures for recognizing cellular subpopulations made it possible to unequivocally define all the steps necessary to determine the main factor of heterogeneity of the analyzed data. A crucial part of the entire process was the feature selection procedures application, which not only allowed for a significant reduction in the dimensionality of the data but also improved the procedures used further by limiting the data set to informative, at the level of cells' differentiation, features. Based on using the HDBSCAN tool for cell clustering, the subsequent analysis stage was associated with a significant increase in computational expenditure and, therefore, closely related to a greater need for time to conduct a specific study stage. It is worth paying particular attention that when tuning the tool's initial parameter values, it is necessary to consider its high sensitivity, even with slight changes in the parameter values. Therefore, an essential step is introducing a specific, consistent measure of the clustering quality. The lack of information about the affiliation of individual cells to appropriate cell subpopulations imposed the use of a measure that was an indicator of good cluster separation due to the only information available, i.e., the counts of cells creating designated subgroups. All the analyzed for subpopulation recognition ex vivo data structures (set A model structure, set B model structure, and set B test structure) achieved very high values of the utilized omega-squared metric equal to 0.97, 0.99, and 0.86, respectively. This means the data was very well separated into clusters of similar diversity using the HDBSCAN tool. The use of information about marker genes of an individual and expected cell subpopulations made it possible to combine clusters with a similar genetic profile in the next stage.

In connection with the conducted series of analyses, it was proved hypothesis 1. that combining feature engineering methods and advanced dimensionality reduction techniques with unsupervised clustering algorithms allows for the efficient identification of white blood cell subtypes in single-cell RNA sequencing data. The analysis workflow and the procedures included enabled the detection of white blood cell subpopulations and the determination of the reason for the observed cellular variability not related to the radiation factor. High internal heterogeneity of the analyzed data was indicated as the source of this variability. The conclusions were supported by applying the detected subpopulations to pre-defined clusters extracted using unsupervised UMAP learning techniques. An almost perfect match was achieved in analyzed ex vivo datasets. For this data type, the visualization of spatially distributed cells and the color coding of the respective subpopulations fully coincided with the clusters. Therefore, the analysis path made it possible to detect several subpopulations of white blood cells, the vast majority of which, as much as 85-90% of the datasets, were T-cells. A monocyte subpopulation was also detected for all subsets of data. For the model structures of datasets A and B, rarer subpopulations such as B cells, granulocytes, and dendritic cells were also detected. For the model structure of dataset B, it was also possible to detect a very small subpopulation of basophils/eosinophils, representing approximately 0.5% of the analyzed white blood

cell data. Such results indicate the high efficiency and sensitivity of the proposed approach for detecting cell subpopulations for data from single-cell sequencing experiments.

9.2 Ex vivo irradiated cells genetic profile

One of this dissertation's most technically valuable aspects is the development of the feature selection procedures workflow. As discussed in the introduction, this phenomenon is one of the most critical stages of high-dimensional data analysis. It enables practical conclusions on the collected data and the analysis results. A workflow based on logistic regression methods was implemented to automate feature selection procedures. This analysis stage should be devoted to a relatively large amount of time due to the number of benefits that can be obtained related to the possibility of correct interpretation of the results. However, it should be noted that the proposed workflow is not a quick-working solution but allows for maintaining all existing relationships between the analyzed features. Moreover, this tool's components cannot only detect significant changes in the genetic profile of irradiated cells. It can also be successfully used for classification purposes, based on a previously created full model, for a set of control and irradiated cells. Implemented workflow enables the analysis to be carried out stably, with a guaranteed choice of the most critical parameters that should be used in case of the need for their manipulation due to the different purposes and specific assumptions of the analysis performed. In the case of estimating model parameters with a list of selected features and in the case of features selection problem, it is possible to manipulate parameters such as *epsilon* (difference of likelihood value between successive estimates of model parameters), *max_iter* (maximum number of iterations of model parameters estimation process) and *alpha* (learning rate for the estimation of model parameters). When performing the testing part, it is possible to change the value of the *pred_thr* parameter, which is responsible for the probability value for the observations' classification, to the positive class. Using this implementation, it is possible to carry out all the necessary steps of model building, testing, and utilizing the constructed model for cell classification of external data sets.

Recognition of the genetic profile for the entire gene pool available in the datasets allowed the detection of a factor worth noting, namely the need to perform a thorough and refined selection of features. The detected genetic profile of the irradiated cells of such a dataset allowed recognition of the radiation response genes. Still, it was contaminated with features not responsible for distinguishing the irradiated cells. Detected genes such as AQP9, CD3D, FYB, LAT, LAMP3, LCK, and TRIB2 were selected at the feature selection stage to detect differences between white blood cell subpopulations present and previously uncontrolled in the dataset. To carry out the appropriate path of data analysis in terms of recognizing the correct genetic profile of cells irradiated in the ex vivo environment, the variability associated with the occurrence of cell subpopulations was removed before, filtering out only the T-cell subpopulation, which constitutes the vast majority of all analyzed cells.

Based on the applied workflow, it was proved hypothesis 2. that the proposed intelligent and stratified algorithm of the training set construction supports the classification system, especially in the case of heterogeneous datasets. This approach made it possible to find out how the control cells differ from irradiated ones without additional disturbing factors. In the gene composition of the model built based on the T-cells normalized counts, almost only the genes described in the literature reports as radiation response genes were found. As many as 18 out of 21 such genes were included in the model. The remaining three genes, HLA-A, CD52, and TRIB2, were associated with the pathways of biological processes responsible mainly for the cellular response to a harmful external factor. The previously performed procedures for detecting cell subpopulations were necessary to remove the cause of the high heterogeneity of the ex vivo data set, which was exhaustively confirmed with the visualization method on separate data subsets using unsupervised learning techniques for this division. As a result, the analysis was conducted in a manner that did not object to UMAP's prior knowledge of the internal differences in the data set. After removing the cause of the heterogeneity of the collection, i.e., the different subpopulations present, and focusing the classifier's attention only on the majority

class of the T-cells subpopulation, it was possible to detect the only cause of differences between cells, i.e., the presence of two cells' types - control and irradiated. The selected set of features fully corresponds to the assumption of recognizing the genetic profile of irradiated cells stated in this dissertation. The constructed genetic profile of irradiated cells consisted of 21 features. Identified radiation-response genes with assigned parameter values are indicated in Table 39.

Table 39. Recognized ex vivo irradiated cells genetic profile.

Intercept	**RPS19P1**	**RPL23AP42**	**BAX**	**DDB2**	**HLA-A**	**RPS27L**	**PHPT1**
-4.47	1.76	-1.44	0.60	0.22	-0.75	0.40	0.32
MYC	**CCNG1**	**FYB**	**CD52**	**CD74**	**TNFSF8**	**THBS1**	**STAT1**
-0.54	0.32	-0.39	-0.39	-0.22	0.10	-0.30	-0.31
TRIB2	**LCK**	**AEN**	**CDKN1A**	**STAT5A**	**CHI3L1**		
-0.32	-0.33	0.06	0.05	-0.29	0.11		

The constructed model of the irradiated cells' genetic profile also allowed high weighted classification accuracy for the test set, equal to 93.83%. Excellent precision, sensitivity, and specificity values of 0.9362, 0.9255, and 0.9487 were also achieved. The percentage of incorrectly classified cells was equal to 6.17% of the test set, which means that only 48 out of 778 observations were not correctly assigned. In addition, comparing the MISO model built based on 21 selected genes and single-factor SISO models showed the advantage of the proposed genetic profile of irradiated cells. Despite the higher complexity, the MISO model achieves noticeably better results considering the $F1_{score}$ measure and the weighted accuracy of cells' classification.

9.3 Logistic regression and neural networks-based workflows

Comparing two machine learning methods regarding feature selection and classification of irradiated and control cells is challenging. Many factors must be taken into account, among which the most important are the degree of complexity of the analysis, computational costs, the time necessary to perform the calculations, and the results achieved in terms of interpretability and correctness.

Considering the complexity of the analysis carried out using logistic regression methods and neural networks, and based on the presented approaches, both workflows cannot be directly compared in this respect. This is because the workflow related to the logistic regression methods was implemented in its entirety for this doctoral dissertation. In contrast, using neural networks was mainly associated with using ready-made and publicly available functions. Moreover, for the logistic regression methods, a more thorough study was carried out related to the selection of features, consisting in randomizing the training set 50 times, which ensured more significant variability and generalization of the problem compared to the approach using neural networks. In addition, in the case of neural networks, a tool supporting the explanation of the model structure was used, which undeniably increased the complexity of this approach. However, apart from differences in the first-line feature selection technique, i.e., wrappers, both approaches were consistent in the further use of the filters method. In both cases, the number of significant features was estimated, and the final model was built on these features.

Another critical factor relates to the computational costs and the time necessary to achieve satisfactory results. In this context, modeling using neural networks is superior to the developed approach based on logistic regression. Obtaining 50 LR-based models for feature selection required significant time and was computationally intensive. In the case of neural networks, only one model was built based on which the significance of individual features was concluded. Thus, the computational costs were relatively small, and the time to obtain results was incomparably shorter. The listed aspects and conclusions do not result from the nature of the operation or the intention of individual machine-learning solutions. This comparison is based on the solutions used in this dissertation and cannot be

generalized and used to draw general conclusions about the effectiveness of both machine learning methods. Regarding the solutions used in this work, the approach to using neural networks is characterized by a much higher level of complexity. In contrast, the approach associated with using logistic regression methods required a much longer calculation time.

However, described factors cannot exceed the importance of the last one, i.e., the interpretability and correctness of the results. Both workflows yielded a genetic profile of irradiated cells in an ex vivo environment. Moreover, as many as eight genes were common to these approaches. This indicates a high consistency in recognizing the appropriate structures of cell changes under the influence of an ionizing radiation agent. Considering the complexity of the identified model of irradiated cells, 21 features were significant in logistic regression methods. In comparison, only ten features were used with neural networks. An additional analysis comparing the classification quality of irradiated cells for both models showed a clear advantage for the model built based on logistic regression methods, enabling almost 94% classification quality and a very high value of the $F1_{score}$ metric at the level above 0.93. It allows better cell classification results and the detection of more radiation response genes. The final model is more complex than in the case of neural networks. However, there is still a significant dimensionality reduction, and the resulting genetic profile does not contain features that are not directly related to the research problem. Moreover, the proposed approach ensured complete control over the data from the feature selection stage through the model tuning stage to the final model use for observations' classification.

10 Discussion

The theses presented in this doctoral dissertation are not trivial aspects regarding the analyses and solutions interpretation. Moreover, the given subject still leaves a lot to maneuver in the context of the possibility of applying improvements and eliminating the need for human intervention in individual stages. Of course, the ability to validate individual results is necessary, but in the developed workflow, there are several individual fragments where improvements and automation could be successfully applied. One of these elements is the procedure related to the recognition of cell subpopulations and, more precisely, the assignment of many individual clusters detected by the HDBSCAN tool to the actual subpopulations of white blood cells. This step requires much human input regarding the time needed to analyze the distribution of all marker genes for each detected cluster. An automated solution based on the counts of these clusters, and more precisely based only on the values of the first, second, and third quartiles of distribution, would enable the accurate and efficient diagnosis of white blood cell subpopulations. Such a solution would significantly improve the work and eliminate the likelihood of human error. Recognition of cell subpopulations is an aspect that certainly positively impacts recognizing the correct genetic profile of irradiated cells. The verification of the genetic profile built based on a heterogeneous set (without separated cell subpopulations) revealed the presence of features that are not related to the research problem but result only from the lack of complete control of the phenomena and relationships occurring in the dataset. The genetic profile of the irradiated cells, built on the set after removing the internal heterogeneity, i.e., after filtering out the collection of T-cells, enabled the detection of those genes that affect the recognition of control and irradiated cells. In the final model, built based on logistic regression methods, no genes responsible for recognizing variability other than those resulting from irradiation of the cell fraction were detected. Comparative analysis performed for two machine learning methods, logistic regression, and neural networks, allowed establishing a specific clamp closing the careful considerations on the theses. This analysis made it clear that despite applying different methods, the radiation response genes are mainly common to these approaches. Moreover, eight genes common to both studies, i.e., AEN, TNFSF8, CCNG1, PHPT1, RPS27L, DDB2, BAX, and RPS19P1, are described in the available and current literature sources as well-known radiation response genes. In addition, despite the apparent advantage of the model built based on logistic regression methods and the classification carried out by this trend, both allow for satisfactory results in recognizing and classifying control and irradiated cells.

There are many possibilities to apply the proposed approaches related to machine learning, the described path of subpopulation recognition, and the use of the final model of the irradiated cells' genetic profile. The implemented machine learning algorithm, based on logistic regression methods, is straightforward to use and allows a very class-oriented reduction of the dimensionality of the features in the set to the most important ones. Moreover, this algorithm consists of a parameter panel that can be adjusted according to the needs of the analysis. Additionally, the advantage of such a self-learning algorithm is the possibility of applying it to two-class problems for biological data and any other available data of interest.

Due to the very high coverage of the recognized genetic profile of ex vivo irradiated cells with current literature reports, it is also possible to use it in further data sets from single-cell sequencing experiments to distinguish control and irradiated cells. It is also worth emphasizing the possibility of using this full model concerning higher doses of absorbed ionizing radiation. However, this approach requires studying how this model copes with possibly more significant differentiation of the control and irradiated class. An interesting approach is to compare the performance of the proposed genetic model of irradiated cells with, as mentioned, doses greater than 1 Gy and also for doses below 1 Gy. A comparative analysis would make it possible to determine the model's universality over the full spectrum of the absorbed radiation dose or to achieve satisfactory results in the specified range of the absorbed doses.

The analysis scheme of recognizing individual cell subpopulations also has a vast application if we consider the diversity of data sets. The developed workflow allows not only effective separate clusters that are biologically differentiated. It also provides for detecting the heterogeneity phenomenon occurring among other data sets. However, this highly complex scheme requires manual interpretation based on the features' count distribution graphs for the detected clusters. This issue must be addressed before introducing the tool to a broader range of conscious and unconscious users. The developed workflow is complicated; it consists of many stages, during which a general knowledge of the analyzed data should be demonstrated and often advanced statistical, mathematical, or programming reasoning. Simplifying the developed series of methods in terms of ease of use would allow the individual stages of the analysis to be packed into a 'box.' It would serve as a framework containing particular methods and allowing for interference by an inexperienced user while eliminating the possibility of making a mistake and drawing hasty and erroneous conclusions. For advanced users, however, it should be possible to have a more significant impact on the operation of the following analysis steps in this 'box'. Such an approach to the modernization of the proposed workflow would solve the undoubted problem of the 'black box' in many publicly available tools and the lack of knowledge about the subsequent stages of analysis and the selection, often arbitrarily considered universal threshold values. An advanced user could fully understand each analysis step and select only those relevant to the research problem. This element is undoubtedly missing in many available and widely used tools for careful data analysis.

Due to the increasing use of tools generating high-dimensional data, the presented approaches are highly universal concerning the diversity and composition of these data. Big and high-dimensional data is produced in industry, science, and medicine. The use of tools to improve the work of analysts, both in terms of time and reducing the possibility of human error, is increasingly necessary and more and more appreciated nowadays. The tools and solutions presented in the doctoral dissertation also have very high development potential, constituting an excellent basis for future considerations.

11 Abstract

Single-cell RNA-sequencing (scRNA-seq) is an increasingly widely used technology to analyze the transcriptome of many single cells. By sequencing the genome of single cells, it is possible to avoid the data generalization problem in sequencing technologies that do not focus on individual cells. As a result of utilizing this technology, high-dimensional data is generated, which requires more and more computing resources to make the proper analysis. The scRNA-seq technology is essential for investigating cell-to-cell heterogeneity in analyzing the impact of specific factors, such as a cellular response to ionizing radiation. Unconscious or conscious exposure to radiation induces changes in cell responses caused by modifications in the expression of many genes regulating cell lives. The analysis of such modifications can reveal genes that are most involved in the radiation response. Such analysis can also demonstrate gene communication pathways that could give insight into what changes occur throughout a complex cellular system. By combining the knowledge about the level of radiation-induced changes and the available sequencing technologies, we can perform appropriate analysis steps that will allow us to learn about the genes that respond to radiation.

This work has two main goals. One is purely biological, while the other is related to engineering. **The biological aim of this study is to search for known and new genes of radiation response based on the data from single-cell RNA-sequencing techniques**. This way, the differences in the gene signature of normal cells and those subjected to ionizing radiation will be determined. **A fundamental goal in engineering is to create an appropriate bioinformatic analysis workflow to partially automate the consecutive steps of working with high-dimensional data from single-cell sequencing experiments.** The main aspects included in the proposed method are based on feature selection procedures and the problem of cell classification itself. It is a considerable challenge, especially considering the very high complexity and dimensionality of the analyzed data, but also the expectations of achieving satisfactory results regarding the quality of the classification of irradiated cells. The expected outcome of the created tool is primarily related to the biological purpose of the research, i.e., to the **recognition of the complete genetic profile of cells irradiated in an ex vivo environment.**

The first work stage focuses on data quality control. For this purpose, two ex vivo samples, technical repetitions of the same experiment, were tested. Several statistical and visualization paths were developed to allow detailed analysis of the quality of both genes and cells. The methodology used, especially the unsupervised classification approach utilized for visualization, allows for drawing an unambiguous conclusion about the significant heterogeneity of the studied data sets. Therefore, attempts were made to determine the cause of such cell heterogeneity using public and own-developed mathematical and statistical methods. Moreover, a list of subpopulation-specific marker genes was also used to designate white blood cell subpopulations. It was proved that the chosen research path determined the cause of the internal heterogeneity in complex data sets related to the occurrence of highly-differentiated cell subtypes. Moreover, as a result of a series of analyses, it was possible to detect frequently occurring subpopulations of this fraction in the quantitative context and rare and small subpopulations of white blood cells. The works' main stage aims to build a classifier based on logistic regression methods. The purpose of the classifier is to distinguish control and ionizing radiation-subjected cells. At this stage of the work, only the T-cells subpopulation was considered as it constituted most of the selected white blood cell subtypes. What is essential, the applied procedure made it possible to remove the substantial heterogeneity of the data set. Next, to standardize the structure of the analyzed data set, there was also performed the data normalization procedure. A feature selection procedure was based on cells and genes prepared this way. For this purpose, an own-implemented workflow was developed, enabling the classification of normal and irradiated cells with adequate measures of classification quality. As a result of using the implemented workflow, a radiation response genes panel was finally obtained. Interestingly, a significant majority of found genes correspond to current literature

reports. While reducing the impact of the heterogeneity in the data set allowed to improve the classification quality to obtain very satisfactory results with the value of the weighted accuracy, based on the hold-out test data set, at above 93%. Additionally, a detailed analysis was made using the neural networks approach to compare logistic regression-based workflow with another well-known method. Another machine-learning analysis workflow was created that is compatible with the stated goal to recognize irradiated cells set out in the dissertation. This approach was primarily aimed at checking and comparing the quality of classifications resulting from using two different feature selection techniques. Using neural networks made it possible to obtain promising results, with a classification quality value of almost 91%. Moreover, such results were achieved in a much shorter time frame, comparing neural networks with a logistic regression-based approach. On the other hand, what is even more critical in undertaking analysis this way, it was also possible to compare the genetic profiles of irradiated cells resulting from the logistic regression and neural networks-based approaches. It occurred that 8 out of 10 genes creating the neural networks-based model are familiar with the logistic regression-based procedure. These well-known genes of radiation response include RPS19P1, BAX, DDB2, RPS27L, PHPT1, CCNG1, TNFSF8, and AEN.

This doctoral dissertation shows that using data derived from a precise and detailed technology, such as scRNA-seq, it is possible to determine the specific gene structure of cells subjected to ionizing radiation. This work also made it possible to compare two machine-learning techniques: logistic regression and neural networks-based approaches. Several bioinformatics methods and different workflows developed can be used in the future as support in medicine, science, and engineering. The developed method for feature selection and irradiated cell classification met the challenges posed in the dissertation with very high efficiency. This research describes exactly the workflow of high-dimensional data analysis from single-cell sequencing experiments, such as the extended quality control, through the recognition of radiation response genes, the determination of the irradiated cells gene signature, classification of white blood cells along with the subpopulations recognition, the comparison of machine learning procedures in terms of high dimensional data analysis and observations' classification, and also the biological interpretation of the results. Therefore this work covers, with a detailed description of the proposed analysis steps and the effects in the form of results, all aspects necessary to achieve the assumed goals, combining them into a logical workflow with appropriate comments and inferences from both the technical and engineering side, and supports these aspects in the form of a biological interpretation.

Sekwencjonowanie RNA pojedynczych komórek (scRNA-seq) jest coraz szerzej stosowaną technologią do analizy transkryptomu wielu pojedynczych komórek. Dzięki zastosowaniu sekwencjonowania genomu pojedynczych komórek można uniknąć problemu generalizacji danych występującego w technologiach sekwencjonowania, które nie koncentrują się na pojedynczych komórkach. W wyniku wykorzystania tej technologii generowane są dane wielowymiarowe, co wymaga coraz większych zasobów obliczeniowych do właściwej analizy. Technologia scRNA-seq jest niezbędna do badania heterogeniczności między komórkami w analizie wpływu określonych czynników, takich jak odpowiedź komórkowa na promieniowanie jonizujące. Zarówno nieświadoma, jak i świadoma ekspozycja na promieniowanie jonizujące wywołuje zmiany w odpowiedziach komórkowych spowodowane modyfikacjami ekspresji wielu genów regulujących życie komórek. Analiza takich modyfikacji może ujawnić geny najbardziej zaangażowane w odpowiedź na promieniowanie. Taka analiza może również wykazać ścieżki komunikacji genów, które mogą dać wgląd w to, jakie zmiany zachodzą w złożonym systemie komórkowym. Integrując wiedzę o stopniu zmian indukowanych promieniowaniem i dostępnymi technologiami sekwencjonowania, możemy przeprowadzić odpowiednie kroki analityczne, które pozwolą nam poznać geny reagujące na promieniowanie.

Przedstawiona rozprawa doktorska ma dwa główne cele. Jeden z nich ma podłoże biologiczne, a drugi związany jest z inżynierią. **Celem biologicznym niniejszej pracy jest poszukiwanie znanych i nowych genów odpowiedzi na promieniowanie w oparciu o dane z technik sekwencjonowania RNA pojedynczych komórek.** W ten sposób zostaną określone różnice w sygnaturze genowej normalnych komórek i tych poddanych działaniu promieniowania jonizującego w środowisku ex vivo. **Podstawowym celem pracy w zakresie inżynierii jest stworzenie odpowiedniego schematu pracy analizy bioinformatycznej, aby częściowo zautomatyzować kolejne etapy pracy z wielowymiarowymi danymi pochodzącymi z eksperymentów sekwencjonowania pojedynczych komórek.** Główne aspekty zawarte w proponowanej metodzie opierają się na procedurach selekcji cech oraz problemie klasyfikacji komórek. Jest to duże wyzwanie, zwłaszcza biorąc pod uwagę bardzo istotną złożoność i wymiarowość analizowanych danych, ale także oczekiwania nastawione na uzyskanie zadowalających wyników w zakresie jakości klasyfikacji komórek napromienionych. Oczekiwany wynik stworzonego narzędzia związany jest przede wszystkim z biologicznym celem badań, tj. **rozpoznaniem pełnego profilu genetycznego komórek napromienionych w środowisku ex vivo.**

Pierwszy etap prac koncentruje się na kontroli jakości danych. W tym celu przetestowano dwie próbki ex vivo, będące technicznymi powtórzeniami tego samego eksperymentu. Opracowano kilka ścieżek statystycznych i wizualizacyjnych, aby umożliwić szczegółową analizę jakości zarówno genów, jak i komórek poddanych analizie. Zastosowana metodologia, a zwłaszcza nienadzorowane podejście klasyfikacyjne zastosowane do celów wizualizacji, pozwala na wyciągnięcie jednoznacznego wniosku o znacznej heterogeniczności badanych zbiorów danych. W związku z tym, podjęto próby ustalenia przyczyny takiej heterogeniczności komórek wykorzystując zarówno ogólnodostępne, jak i opracowane na potrzeby niniejszej rozprawy metody matematyczno-statystyczne. Ponadto, do rozpoznania subpopulacji komórek białych krwinek, wykorzystano również listę genów markerowych specyficznych dla określonych subpopulacji. Wykazano, że wybrana ścieżka badawcza pozwoliła na określenie przyczyny wewnętrznej heterogeniczności, w złożonych zbiorach danych, związanej z występowaniem wysoko zróżnicowanych podtypów komórek. Co więcej, w wyniku przeprowadzonych serii analiz udało się wykryć często występujące subpopulacje tej frakcji, w kontekście ilościowym, oraz rzadkie i małe subpopulacje białych krwinek. Główny etap prac ma na celu zbudowanie klasyfikatora opartego o metody regresji logistycznej. Zadaniem klasyfikatora jest rozróżnienie komórek kontrolnych i poddanych działaniu promieniowania jonizującego. Na tym etapie pracy uwzględniono jedynie subpopulację limfocytów T, gdyż stanowiła ona większość rozpoznanych podtypów analizowanych komórek białych krwinek. Co istotne, zastosowana procedura pozwoliła na usunięcie czynnika odpowiedzialnego za występowanie wykrytej heterogeniczności zbioru danych. Następnie, w celu ujednolicenia struktury analizowanego zbioru danych,

przeprowadzono procedurę normalizacji. Kolejny etap selekcji cech został oparty na zbiorze komórek i genów przygotowany w przedstawiony sposób. Celem selekcji cech, samodzielnie opracowano i zaimplementowano schemat analizy, umożliwiający klasyfikację komórek normalnych i napromieniowanych, z wykorzystaniem odpowiednich miar jakości klasyfikacji. W wyniku zastosowania zaimplementowanego algorytmu uzyskano ostatecznie panel genów odpowiedzi na napromienienie. Co istotne, znaczna większość rozpoznanych genów odpowiedzi radiacyjnej odpowiada aktualnym doniesieniom literaturowym. Zmniejszenie wpływu heterogeniczności w zbiorze danych pozwoliło na poprawę jakości klasyfikacji i uzyskanie bardzo zadowalających wyników z wartością ważonej jakości klasyfikacji, opartej na testowym zbiorze danych, na poziomie powyżej 93%. Dodatkowo przeprowadzono szczegółową analizę z wykorzystaniem podejścia sieci neuronowych w celu porównania schematu pracy opartego o metody regresji logistycznej z inną, dobrze znaną metodą uczenia maszynowego. Utworzono drugi schemat analizy, który jest spójny z określonym celem rozprawy, czyli rozpoznaniem komórek napromienionych. Podejście to miało na celu przede wszystkim sprawdzenie i porównanie jakości klasyfikacji wynikających z zastosowania dwóch różnych technik selekcji cech. Wykorzystanie sieci neuronowych pozwoliło uzyskać obiecujące wyniki, z wartością ważonej jakości klasyfikacji na poziomie prawie 91%. Co więcej, takie wyniki uzyskano w znacznie krótszym czasie, porównując sieci neuronowe z podejściem opartym o metody regresji logistycznej. Z drugiej strony, co jeszcze ważniejsze przy przeprowadzaniu analiz zgodnie z zaproponowanym schematem analizy, możliwe było również porównanie profili genetycznych komórek napromienionych, rozpoznanych w wyniku zastosowania metod regresji logistycznej oraz sieci neuronowych. Okazało się, że 8 na 10 genów tworzących model oparty o sieci neuronowe jest spójnych z modelem opartym o regresję logistyczną. Te dobrze znane geny odpowiedzi na promieniowanie obejmują RPS19P1, BAX, DDB2, RPS27L, PHPT1, CCNG1, TNFSF8 i AEN.

Niniejsza rozprawa doktorska pokazuje, że wykorzystując dane pochodzące z precyzyjnej i szczegółowej technologii, takiej jak scRNA-seq, można określić specyficzną strukturę genów dla komórek poddanych działaniu promieniowania jonizującego. Przeprowadzone prace umożliwiły również porównanie dwóch technik uczenia maszynowego w kontekście selekcji cech. Kilka opracowanych metod bioinformatycznych, a przede wszystkim zaproponowany schemat analizy, mogą być w przyszłości wykorzystane jako wsparcie w medycynie, nauce i inżynierii. Opracowana metoda selekcji cech i klasyfikacji komórek napromienionych sprostała wyzwaniom postawionym w rozprawie z bardzo wysoką skutecznością. Badania te dokładnie opisują przebieg analizy danych wysokowymiarowych, pochodzących z eksperymentów sekwencjonowania pojedynczych komórek, takich jak: rozszerzona kontrola jakości, rozpoznanie genów odpowiedzi na promieniowanie, określenie sygnatury genowej komórek napromienionych, klasyfikację białych krwinek wraz z rozpoznawaniem określonych subpopulacji komórkowych, porównanie procedur uczenia maszynowego pod kątem analizy danych wysokowymiarowych i klasyfikacji obserwacji, a także biologiczną interpretację wyników. Niniejsza praca obejmuje, wraz ze szczegółowym opisem proponowanych etapów analizy oraz efektów w postaci wyników, wszystkie aspekty niezbędne do osiągnięcia założonych celów, łącząc je w logiczny schemat pracy wraz z odpowiednimi komentarzami i wnioskami zarówno od strony technicznej, inżynieryjnej, jak i wsparcia tych aspektów w postaci interpretacji biologicznej.

Acknowledgments

This work has been supported by European Union under the European Social Fund grant AIDA – POWR.03.02.00-00-I029.

References

[1] N. I. Zakariya and M. T. E. Kahn, "Benefits and biological effects of ionizing radiation," *Scholars academic journal of biosciences, 2(9),* pp. 583-591, (2014).

[2] E. Shapiro, T. Biezuner and S. Linnarsson, "Single-cell sequencing-based technologies will revolutionize whole-organism science," *Nature Reviews Genetics 14.9,* pp. 618-630, 2013.

[3] B. Munsky, G. Neuert and A. Van Oudenaarden, "Using gene expression noise to understand gene regulation," *Science 336.6078,* pp. 183-187, 2012.

[4] A. Kulkarni, A. G. Anderson, D. P. Merullo and G. Konopka, "Beyond bulk: a review of single cell transcriptomics methodologies and applications," *Current opinion in biotechnology, 58,* pp. 129-136, 2019.

[5] A. Raj and A. Van Oudenaarden, "Nature, nurture, or chance: stochastic gene expression and its consequences," *Cell 135.2 ,* pp. 216-226, 2008.

[6] O. Stegle, S. A. Teichmann and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nature Reviews Genetics 16.3,* pp. 133-145, 2015.

[7] "Partek Flow," [Online]. Available: https://www.partek.com/partek-flow/. [Accessed 09 09 2022].

[8] C. Megill, B. Martin, C. Weaver, S. Bell, L. Prins, S. Badajoz and A. Carr, "Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices," *bioRxiv,* 2021.

[9] "ROSALIND," [Online]. Available: https://www.rosalind.bio/single-cell-rna-analysis. [Accessed 09 09 2022].

[10] "Cellenics," Center For Computational Biomedicine, Harvard Medical School, [Online]. Available: https://computationalbiomed.hms.harvard.edu/tools-technologies/cellenics/. [Accessed 09 09 2022].

[11] N. L. Bray, H. Pimentel, P. Melsted and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nature biotechnology, 34(5),* pp. 525-527, 2016.

[12] B. Kaminow, D. Yunusov and A. Dobin, "STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data," *Biorxiv,* 2021.

[13] D. G. Bunis, J. Andrews, G. K. Fragiadakis, T. D. Burt and M. Sirota, "dittoSeq: universal user-friendly single-cell and bulk RNA sequencing visualization toolkit," *Bioinformatics, 36(22-23),* pp. 5535-5536, 2021.

[14] "Radiation Sources and Doses," EPA United States Environmental Protection Agency, [Online]. Available: https://www.epa.gov/radiation/radiation-sources-and-doses. [Accessed 28 06 2022].

[15] A. Ruano-Ravina and R. Wakeford, "The increasing exposure of the global population to ionizing radiation," *Epidemiology 31.2,* pp. 155-159, 2020.

[16] P. Curie, "Radioactive substances, especially radium," Nobel lecture 6, 1905.

[17] "Radiation from the Earth (Terrestrial Radiation)," Centers for Disease Control and Prevention, [Online]. Available: https://www.cdc.gov/nceh/radiation/terrestrial.html. [Accessed 28 06 2022].

[18] P. R. J. Burch and F. W. Spiers, "Radioactivity of the human being," *Science, 120(3122),* pp. 719-720, 1954.

[19] "Man-Made Sources," United States Nuclear Regulatory Commission U.S.NRC, [Online]. Available: https://www.nrc.gov/about-nrc/radiation/around-us/sources/man-made-sources.html. [Accessed 28 06 2022].

[20] J. Thurston, "NCRP Report No. 160: ionizing radiation exposure of the population of the United States," 2010.

[21] "Health Effects of Radiation: Health Effects Depend on the Dose," Centers for Disease Control and Prevention CDC, [Online]. Available: https://www.cdc.gov/nceh/radiation/dose.html. [Accessed 30 06 2022].

[22] E. H. Donnelly, J. B. Nemhauser, J. M. Smith, Z. N. Kazzi, E. B. Farfan, A. S. Chang and S. F. Naeem, "Acute radiation syndrome: assessment and management," *Southern medical journal, 103(6),* p. 541, 2010.

[23] J. Valentin, "The 2007 recommendations of the international commission on radiological protection," *International Commission on Radiological Protection: Elsevier,* 2008.

[24] A. Haque, J. Engel, S. A. Teichmann and T. Lönnberg, "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications," *Genome medicine, 9(1),* pp. 1-12, 2017.

[25] A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney and N. Yosef, "Science forum: the human cell atlas," *elife, 6, e27041,* 2017.

[26] J. K. De Kanter, P. Lijnzaad, T. Candelli, T. Margaritis and F. C. Holstege, "CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing," *Nucleic acids research, 47(16),* pp. e95-e95, 2019.

[27] A. Kriegstein, A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat and P. Chen, "Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex," *Nature biotechnology, 32(10),* pp. 1053-1058, 2014.

[28] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky and I. Amit, "Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types," *Science, 343(6172),* pp. 776-779, 2014.

[29] V. Proserpio and B. Mahata, "Single-cell technologies to study the immune system," *Immunology, 147(2),* pp. 133-140, 2016.

[30] "White Blood Cell Count (WBC) and Differential," [Online]. Available: https://www.rnceus.com/cbc/cbcwbc.html. [Accessed 09 09 2022].

[31] "Blood Differential Test," healthline, [Online]. Available: https://www.healthline.com/health/blood-differential. [Accessed 09 09 2022].

[32] R. C. Wilkins, D. Wilkinson, H. P. Maharaj, P. V. Bellier, M. B. Cybulski and J. R. N. McLean, "Differential apoptotic response to ionizing radiation in subpopulations of human white blood cells," *Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 513(1-2),* pp. 27-36, 2002.

[33] J. R. Quinlan, "Induction of decision trees," in *Machine learning 1.1*, 1986, pp. 81-106.

[34] B. Krose and P. V. D. Smagt, An introduction to neural networks, 2011.

[35] M. Mitchell, An introduction to genetic algorithms, MIT press, 1998.

[36] O. Sutton, "Introduction to k nearest neighbour classification and condensed nearest neighbour data reduction," *University lectures, University of Leicester 1 ,* 2012.

[37] H. A. Park, "An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain," *Journal of Korean Academy of Nursing 43.2,* pp. 154-164, 2013.

[38] K. Sieradzka and J. Polańska, "Feature selection methods for classification purposes," *Recent Advances in Computational Oncology and Personalized Medicine Volume 2: The challenges of the future, Publishing House of the Silesian University of Technology,* pp. 169-189, 2022.

[39] A. Jović, K. Brkić and N. Bogunović, "A review of feature selection methods with applications.," *In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO),* pp. 1200-1205, 2015.

[40] G. Chandrashekar and F. Sahin, "A survey on feature selection methods.," *Computers & Electrical Engineering, 40(1),* pp. 16-28, 2014.

[41] J. Tang, S. Alelyani and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications,* p. 37, 2014.

[42] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," *In Icml (Vol. 1),* pp. 74-81, 2001.

[43] M. Mera-Gaona, D. M. López, R. Vargas-Canas and U. Neumann, " Framework for the ensemble of feature selection methods," *Applied Sciences, 11(17),* p. 8122, 2021.

[44] A. J. Ferreira and M. A. Figueiredo, "Efficient feature selection filters for high-dimensional data," *Pattern recognition letters, 33(13),* pp. 1794-1804, 2012.

[45] S. Solorio-Fernández, J. A. Carrasco-Ochoa and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," *Artificial Intelligence Review, 53(2),* pp. 907-948, 2020.

[46] H. H. Hsu, C. W. Hsieh and M. D. Lu, "Hybrid feature selection by combining filters and wrappers," *Expert Systems with Applications, 38(7),* pp. 8144-8150, 2011.

[47] P. Křížek, J. Kittler and V. Hlaváč, "Improving stability of feature selection methods. In International Conference on Computer Analysis of Images and Patterns," *Springer, Berlin, Heidelberg,* pp. 929-936, 2007.

[48] E. Y. Shum, E. M. Walczak, C. Chang and H. Christina Fan, "Quantitation of mRNA transcripts and proteins using the BD Rhapsody™ single-cell analysis system," *Single molecule and single cell sequencing,* pp. 63-79, 2019.

[49] L. McInnes, J. Healy and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426,* 2018.

[50] A. Coenen and A. Pearce, "A deeper dive into UMAP theory," [Online]. Available: https://pair-code.github.io/understanding-umap/supplement.html. [Accessed 17 09 2022].

[51] A. Coenen and A. Pearce, "Understanding UMAP," [Online]. Available: https://pair-code.github.io/understanding-umap/. [Accessed 17 09 2022].

[52] R. J. Campello, D. Moulavi and J. Sander, "Density-based clustering based on hierarchical density estimates.," in *Pacific-Asia conference on knowledge discovery and data mining (pp. 160-172)*, Springer, Berlin, Heidelberg, 2013.

[53] L. McInnes, J. Healy and S. Astels, "How HDBSCAN Works," [Online]. Available: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html. [Accessed 19 09 2022].

[54] L. McInnes, J. Healy and S. Astels, "How Soft Clustering for HDBSCAN Works," [Online]. Available: https://hdbscan.readthedocs.io/en/latest/soft_clustering_explanation.html. [Accessed 19 09 2022].

[55] D. Jurafsky and J. H. Martin, "Logistic Regression," in *Speech and Language Processing, Chapter 5*, Draft of January 7, 2023.

[56] "tf.keras.Sequential," [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/Sequential. [Accessed 23 01 2023].

[57] "Activation Functions in Neural Networks [12 Types & Use Cases]," [Online]. Available: https://www.v7labs.com/blog/neural-networks-activation-functions. [Accessed 23 01 2023].

[58] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *Journal of pharmaceutical and biomedical analysis 22.5 (2000),* pp. 717-727.

[59] B. D. Ripley, Pattern recognition and neural networks, Cambridge university press, 2007.

[60] S. Bock and M. Weiß, "A proof of local convergence for the Adam optimizer," *In 2019 International Joint Conference on Neural Networks (IJCNN),* pp. 1-8, 2019.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," p. arXiv preprint arXiv:1412.6980, 2014.

[62] Y. Lin, S. Ghazanfar, D. Strbenac, A. Wang, E. Patrick, D. M. Lin and P. Yang, "Evaluating stably expressed genes in single cells," *GigaScience, 8(9), giz106,* 2019.

[63] E. Eisenberg and E. Y. Levanon, "Human housekeeping genes, revisited," *TRENDS in Genetics, 29(10),* pp. 569-574, 2013.

[64] H. J. De Jonge, R. S. Fehrmann, E. S. de Bont, R. M. Hofstra, F. Gerbens, W. A. Kamps and A. ter Elst, "Evidence based selection of housekeeping genes," *PloS one, 2(9), e898,* 2007.

[65] T. Smith, A. Heger and I. Sudbery, " UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy," *Genome research 27.3 ,* pp. 491-499., 2017.

[66] C. Leys, C. Ley, O. Klein, P. Bernard and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of experimental social psychology 49.4,* pp. 764-766, 2013.

[67] V. Vivcharuk, J. Baardsnes, C. Deprez, T. Sulea, M. Jaramillo, C. R. Corbeil, A. Mullick, J. Magoon, A. Marcil, Y. Durocher, M. D. O'Connor-McCourt and E. O. Purisima, "Assisted design of antibody and protein therapeutics (ADAPT)," *PLoS One, 12(7), e0181490,* 2017.

[68] P. J. Rousseeuw and A. M. Leroy, "Robust regression and outlier detection," John wiley & sons, 2005.

[69] "GeneCards: The Human Gene Database," [Online]. Available: https://www.genecards.org/. [Accessed 31 08 2022].

[70] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research, 15(1),* pp. 1929-1958, 2014.

[71] N. Srivastava, "Improving neural networks with dropout. University of Toronto, 182(566)," p. 7, 2013.

[72] J. Brownlee, "What is the Difference Between a Batch and an Epoch in a Neural Network," *Machine Learning Mastery,* p. 20, 2018.

[73] "The Sequential model," [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/Sequential. [Accessed 10 02 2023].

[74] "Welcome to the SHAP documentation," [Online]. Available: https://shap.readthedocs.io/en/latest/. [Accessed 10 02 2023].

[75] W. E. Marcílio and D. M. Eler, "From explanations to feature selection: assessing SHAP values as feature selection mechanism," in *33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*, 340-347, 2020.

[76] Y. G. Lee, J. Y. Oh, D. Kim and G. Kim, "SHAP Value-Based Feature Importance Analysis for Short-Term Load Forecasting," *Journal of Electrical Engineering & Technology,* pp. 1-10, 2022.

[77] L. S. Shapley, "A value for n-person games," 1953.

[78] Y. Narahari, "The Shapley Value," 2012.

[79] "Radiation Basics," United States Environmental Protection Agency EPA, [Online]. Available: https://www.epa.gov/radiation/radiation-basics. [Accessed 29 06 2022].

[80] L.-G. W. Y. H. Q.-Y. H. Guangchuang Yu, "clusterProfiler: an R package for comparing biological themes among gene clusters," *OMICS: A Journal of Integrative Biology 16(5),* pp. 284-287, 2012.

[81] G. Yu, L. G. Wang, Y. Han and Q. Y. He, "clusterProfiler: an R package for comparing biological themes among gene clusters," *OMICS: A Journal of Integrative Biology, 16(5),* pp. 284-287, 2012.

[82] H. Budworth, A. M. Snijders, F. Marchetti, B. Mannion, S. Bhatnagar, E. Kwoh and A. J. Wyrobek, "DNA repair and cell cycle biomarkers of radiation exposure and inflammation stress in human blood.," *PloS one, 7(11), e48619,* 2012.

[83] L. Cruz-Garcia, G. O'Brien, B. Sipos, S. Mayes, M. I. Love, D. J. Turner and C. Badie, "Generation of a transcriptional radiation exposure signature in human blood using long-read nanopore sequencing," *Radiation research, 193(2),* pp. 143-154, 2020.

[84] M. Moreno-Villanueva, Y. Zhang, A. Feiveson, B. Mistretta, Y. Pan, S. Chatterjee and H. Wu, "Single-cell RNA-sequencing identifies activation of TP53 and STAT1 pathways in human T lymphocyte subpopulations in response to ex vivo radiation exposure," *International journal of molecular sciences, 20(9),* p. 2316, 2019.

[85] H. Kaatsch, B. V. Becker, S. Schüle, P. Ostheim, K. Nestler, J. Jakobi and R. Ullmann, "Gene expression changes and DNA damage after ex vivo exposure of peripheral blood cells to various CT photon spectra," *Scientific Reports, 11(1),* pp. 1-9, 2021.

[86] E. Macaeva, M. Mysara, W. H. De Vos, S. Baatout and R. Quintens, "Gene expression-based biodosimetry for radiological incidents: Assessment of dose and time after radiation exposure," *International Journal of Radiation Biology, 95(1),* pp. 64-75, 2019.

[87] A. Tichy, S. Kabacik, G. O'Brien, J. Pejchal, Z. Sinkorova, A. Kmochova and C. Badie, "The first in vivo multiparametric comparison of different radiation exposure biomarkers in human blood," *PLoS One, 13(2), e0193412,* 2018.

[88] S. Paul and S. A. Amundson, "Gene expression signatures of radiation exposure in peripheral white blood cells of smokers and non-smokers," *International journal of radiation biology, 87(8),* pp. 791-801, 2011.

[89] C. Sakakura, K. Miyagawa, K. I. Fukuda, S. Nakashima, T. Yoshikawa, S. Kin and Y. Ito, "Frequent silencing of RUNX3 in esophageal squamous cell carcinomas is associated with radioresistance and poor prognosis," *Oncogene, 26(40) ,* pp. 5927-5938, 2007.

[90] T. Tanaka, T. Bai, K. Yukawa, T. Utsunomiya and N. Umesaki, "Reduced radiosensitivity and increased CD40 expression in cyclophosphamide-resistant subclones established from human cervical squamous cell carcinoma cells," *Oncology reports, 14(4),* pp. 941-948, 2005.

[91] S. A. Ghandhi, L. Smilenov, I. Shuryak, M. Pujol-Canadell and S. A. Amundson, "Discordant gene responses to radiation in humans and mice and the role of hematopoietically humanized mice in the search for radiation biomarkers," *Scientific reports, 9(1),* pp. 1-13, 2019.

[92] K. A. Pilones, M. Charpentier, E. Garcia-Martinez, C. Daviaud, J. Kraynak, J. Aryankalayil and S. Demaria, "Radiotherapy Cooperates with IL15 to Induce Antitumor Immune ResponsesRadiotherapy and IL15 Induce Antitumor Immune Responses," *Cancer immunology research, 8(8),* pp. 1054-1063, 2020.

[93] H. Lyng, K. S. Landsverk, E. Kristiansen, P. M. DeAngelis, A. H. Ree, O. Myklebost and T. Stokke, "Response of malignant B lymphocytes to ionizing radiation: gene expression and genotype," *International journal of cancer, 115(6),* pp. 935-942, 2005.

[94] T. Zhou, J. W. Chou, D. A. Simpson, Y. Zhou, T. E. Mullen, M. Medeiros and W. K. Kaufmann, "Profiles of global gene expression in ionizing-radiation–damaged human diploid fibroblasts reveal synchronization behind the G1 checkpoint in a G0-like state of quiescence," *Environmental health perspectives, 114(4),* pp. 553-559, 2006.

[95] J. A. Siegel, D. Yeldell, D. M. Goldenberg, M. G. Stabin, R. B. Sparks, R. M. Sharkey and R. D. Blumenthal, "Red marrow radiation dose adjustment using plasma FLT3-L cytokine levels: improved correlations between hematologic toxicity and bone marrow dose for radioimmunotherapy patients," *Journal of nuclear medicine, 44(1),* pp. 67-76, 2003.

[96] M. Iwakawa, T. Ohno, K. Imadome, M. Nakawatari, K. I. Ishikawa, M. Sakai and T. Imai, "The radiation-induced cell-death signaling pathway is activated by concurrent use of cisplatin in sequential biopsy specimens from patients with cervical cancer," *Cancer biology & therapy, 6(6),* pp. 905-911, 2007.

[97] L. Cruz-Garcia, G. O'brien, E. Donovan, L. Gothard, S. Boyle, A. Laval and C. Badie, "Influence of confounding factors on radiation dose estimation in in vivo validated transcriptional biomarkers," *Health physics, 115(1),* p. 90, 2018.

[98] C. Girardi, C. De Pittà, S. Casara, G. Sales, G. Lanfranchi, L. Celotti and M. Mognato, "Analysis of miRNA and mRNA expression profiles highlights alterations in ionizing radiation response of human lymphocytes under modeled microgravity," *PLoS One, 7(2), e31293,* 2012.

[99] Y. Feng, Z. Wang, N. Yang, S. Liu, J. Yan, J. Song and Y. Zhang, "Identification of biomarkers for cervical cancer radiotherapy resistance based on RNA sequencing data," *Frontiers in Cell and Developmental Biology, 9,* p. 724172, 2021.

[100] M. S. Islam, M. E. Stemig, Y. Takahashi and S. K. Hui, "Radiation response of mesenchymal stem cells derived from bone marrow and human pluripotent stem cells," *Journal of radiation research, 56(2),* pp. 269-277, 2015.

[101] E. Kis, T. Szatmári, M. Keszei, R. Farkas, O. Ésik, K. Lumniczky and G. Sáfrány, "Microarray analysis of radiation response genes in primary human fibroblasts," *International Journal of Radiation Oncology* Biology* Physics, 66(5),* pp. 1506-1514, 2006.

[102] Z. C. Fu, F. M. Wang and J. M. Cai, "Gene expression changes in residual advanced cervical cancer after radiotherapy: indicators of poor prognosis and radioresistance?," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research, 21,* p. 1276, 2015.

[103] M. A. Chaudhry, "Analysis of gene expression in normal and cancer cells exposed to [gamma]-radiation," *BioMed Research International,* 2008.

[104] A. C. Wilkins, E. C. Patin, K. J. Harrington and A. A. Melcher, "The immunological consequences of radiation-induced DNA damage," *The Journal of Pathology, 247(5),* pp. 606-614, 2019.

[105] Y. C. Kim, M. Barshishat-Kupper, E. A. McCart, G. P. Mueller and R. M. Day, "Bone marrow protein oxidation in response to ionizing radiation in C57BL/6J mice," *Proteomes, 2(3),* pp. 291-302, 2014.

[106] V. Golubovskaya and L. Wu, "Different subsets of T cells, memory, effector functions, and CAR-T immunotherapy," *Cancers, 8(3),* p. 36, 2016.

[107] J. Valentin, J. D. Boice Jr, R. H. Clarke, C. Cousins, A. J. Gonzalez, J. Lee and Z. Q. Pan, "Published on behalf of the International Commission on Radiological Protection," 2007.

[108] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research, 3(Mar),* pp. 1157-1182, 2003.

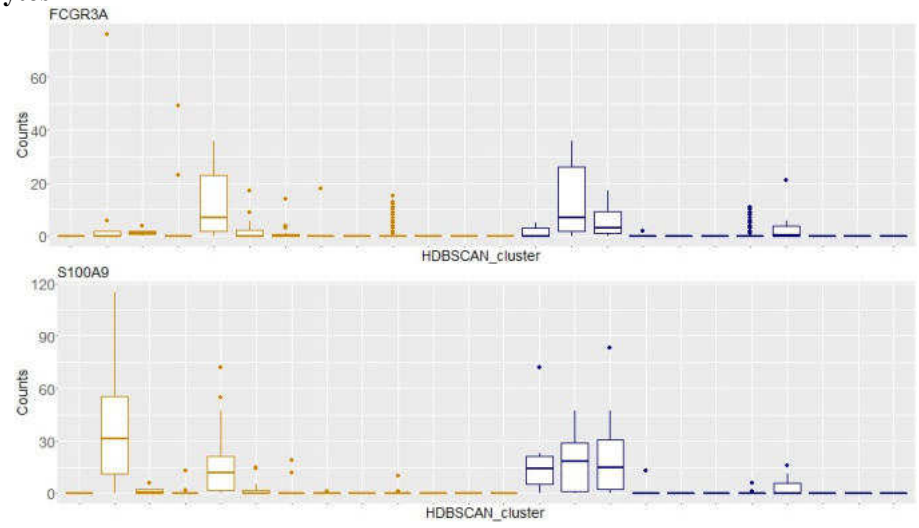
Additional materials

Recognition of cell subpopulations based on ex vivo experiments

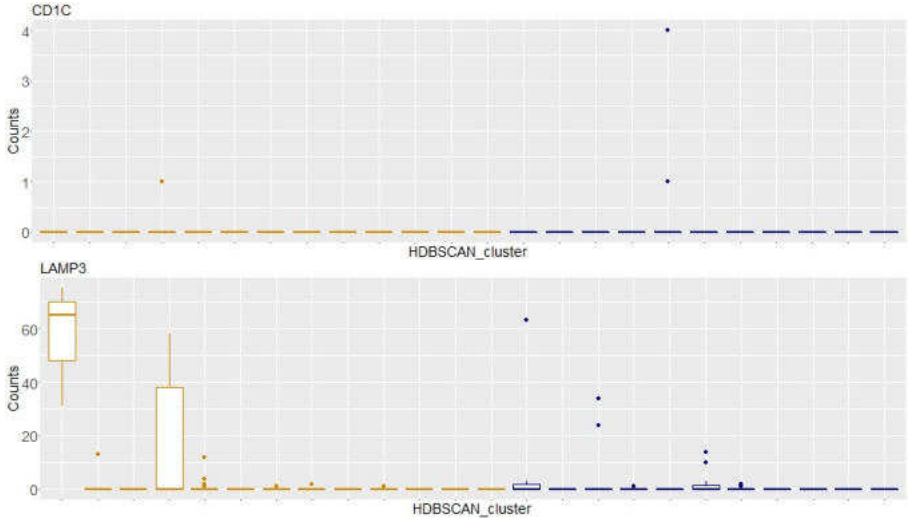
The drawings in this part of the work were used to identify appropriate white blood cell subpopulations of ex vivo datasets. The subsections have been additionally divided according to the subpopulation type. The following graphs of the counts' distribution for clusters separated using the HDBSCAN tool are presented. Clusters designated for **control** cells are marked in goldenrod, and clusters for **irradiated** cells are marked in navy blue.

Set A model structure data

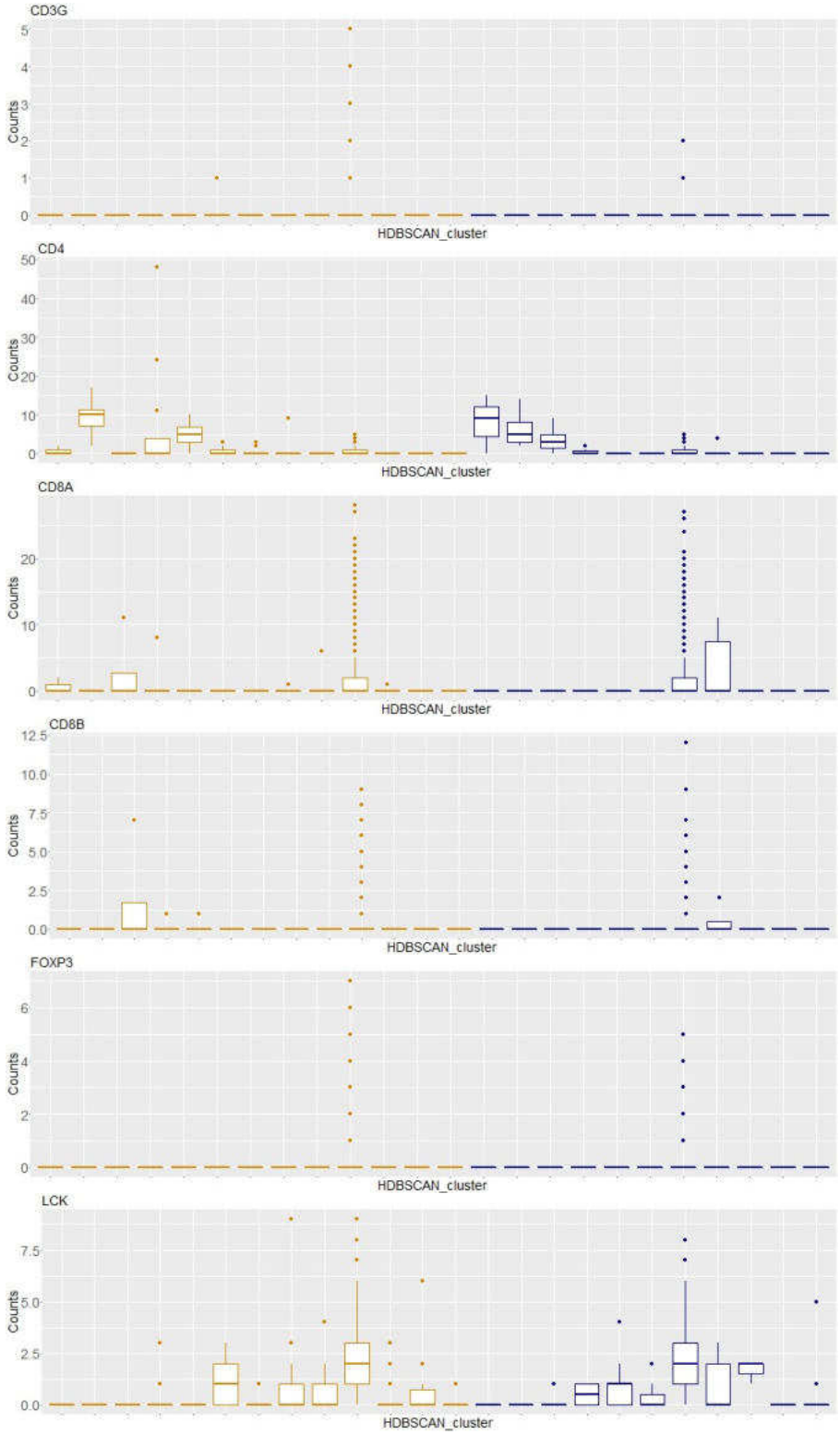
Monocytes



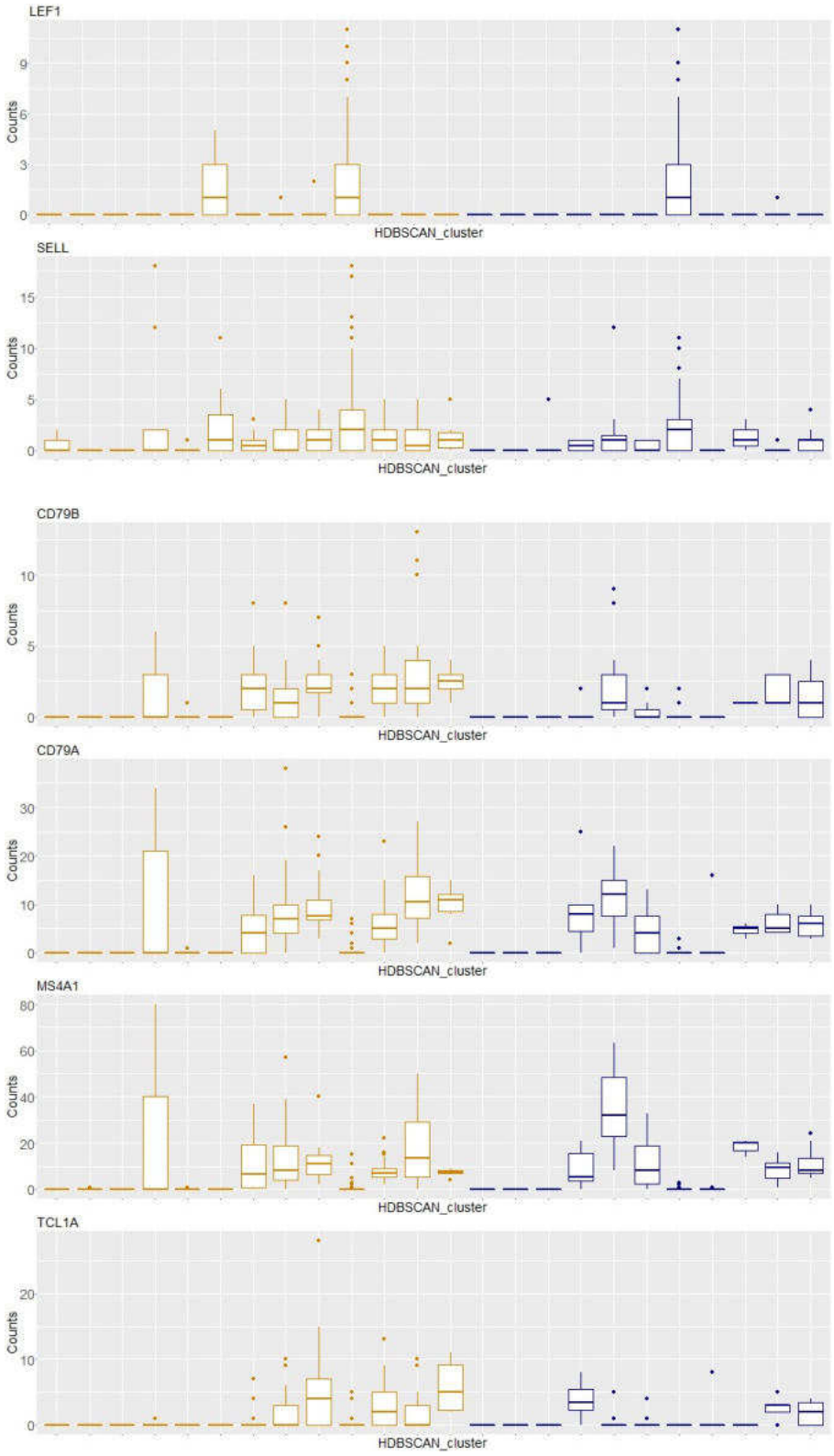
Dendritic cells



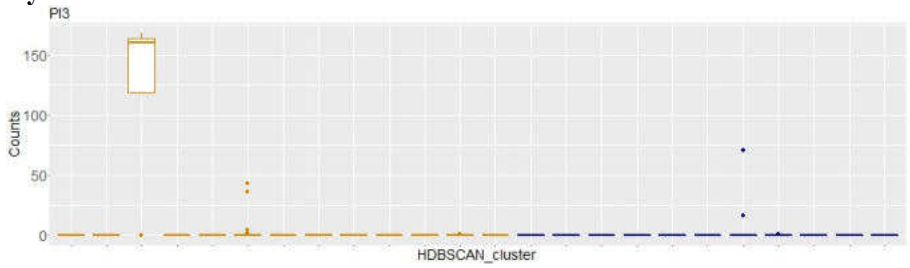
T cells


CD3G
Counts
HDBSCAN_cluster
CD4
Counts
HDBSCAN_cluster
CD8A
Counts
HDBSCAN_cluster
CD8B
Counts
HDBSCAN_cluster
FOXP3
Counts
HDBSCAN_cluster
LCK
Counts
HDBSCAN_cluster

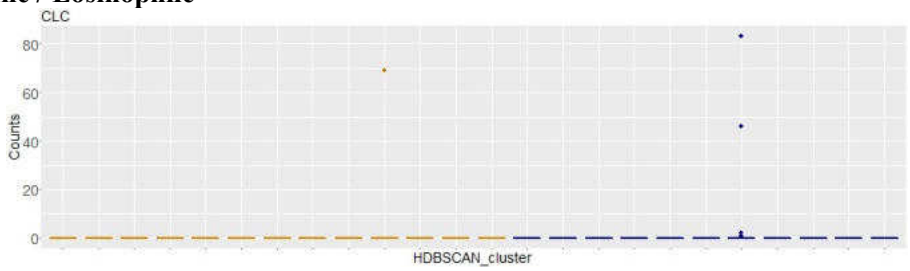
B cells


LEF1
Counts
HDBSCAN_cluster
SELL
Counts
HDBSCAN_cluster
CD79B
Counts
HDBSCAN_cluster
CD79A
Counts
HDBSCAN_cluster
MS4A1
Counts
HDBSCAN_cluster
TCL1A
Counts
HDBSCAN_cluster

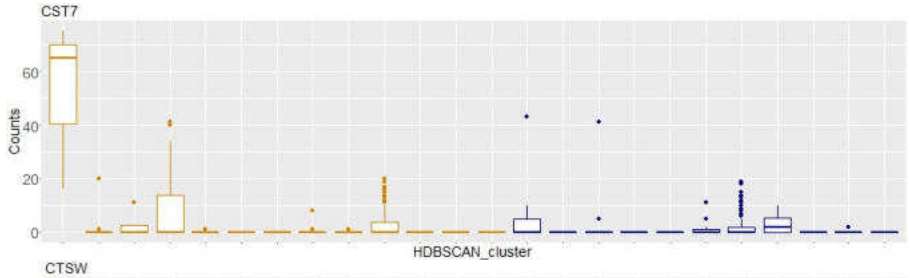
Granulocytes

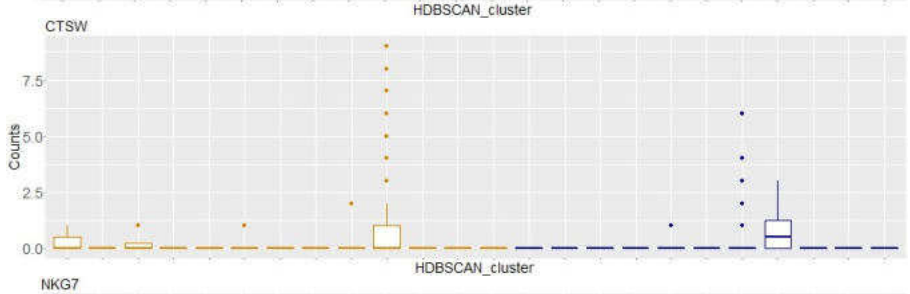


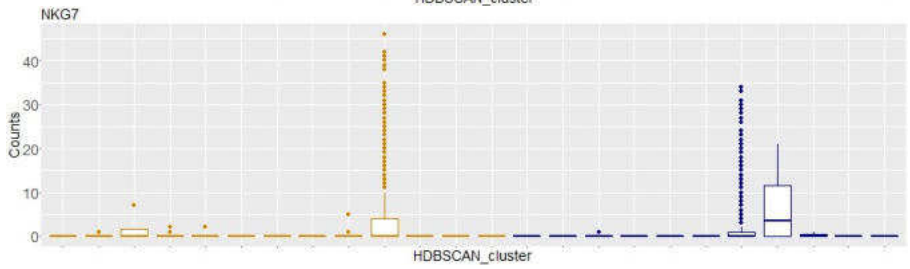
Basophile / Eosinophile



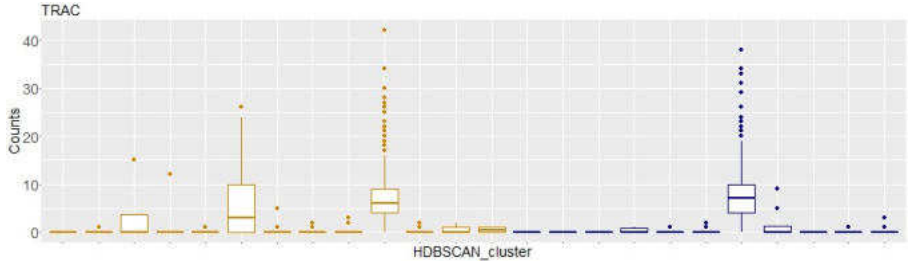
NK cells

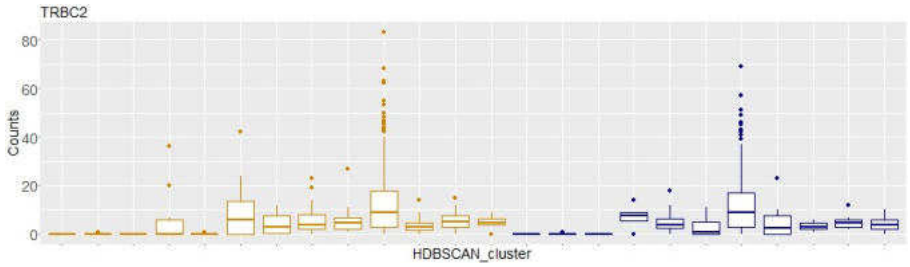




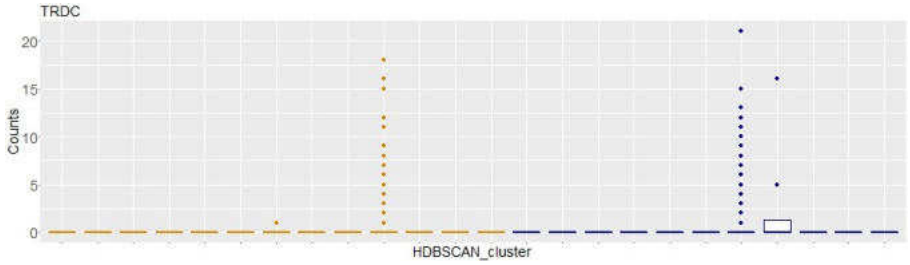


Alpha / Beta T cells



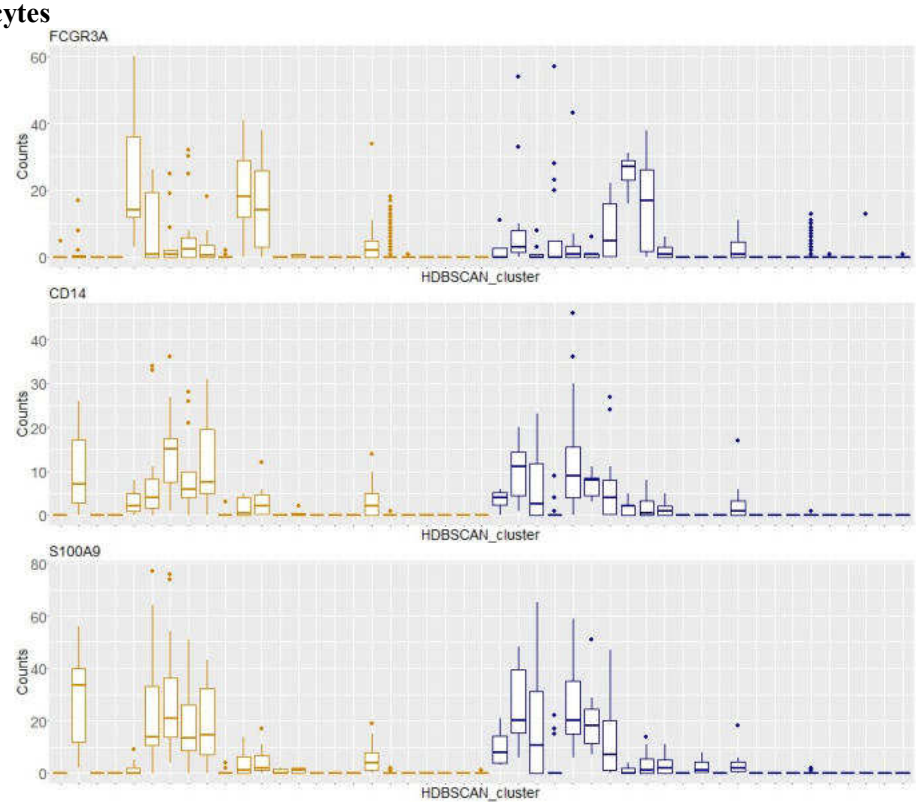


Gamma / Delta T cells

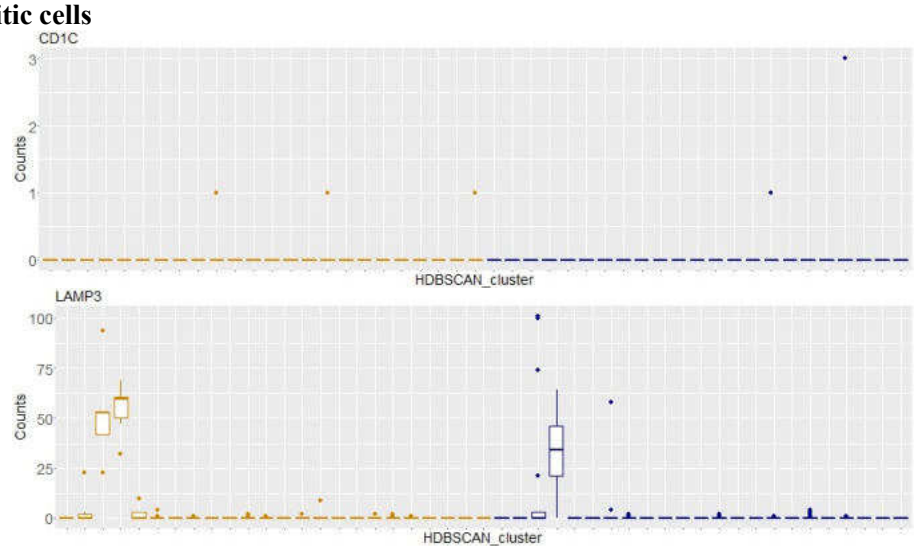


Set B model structure data

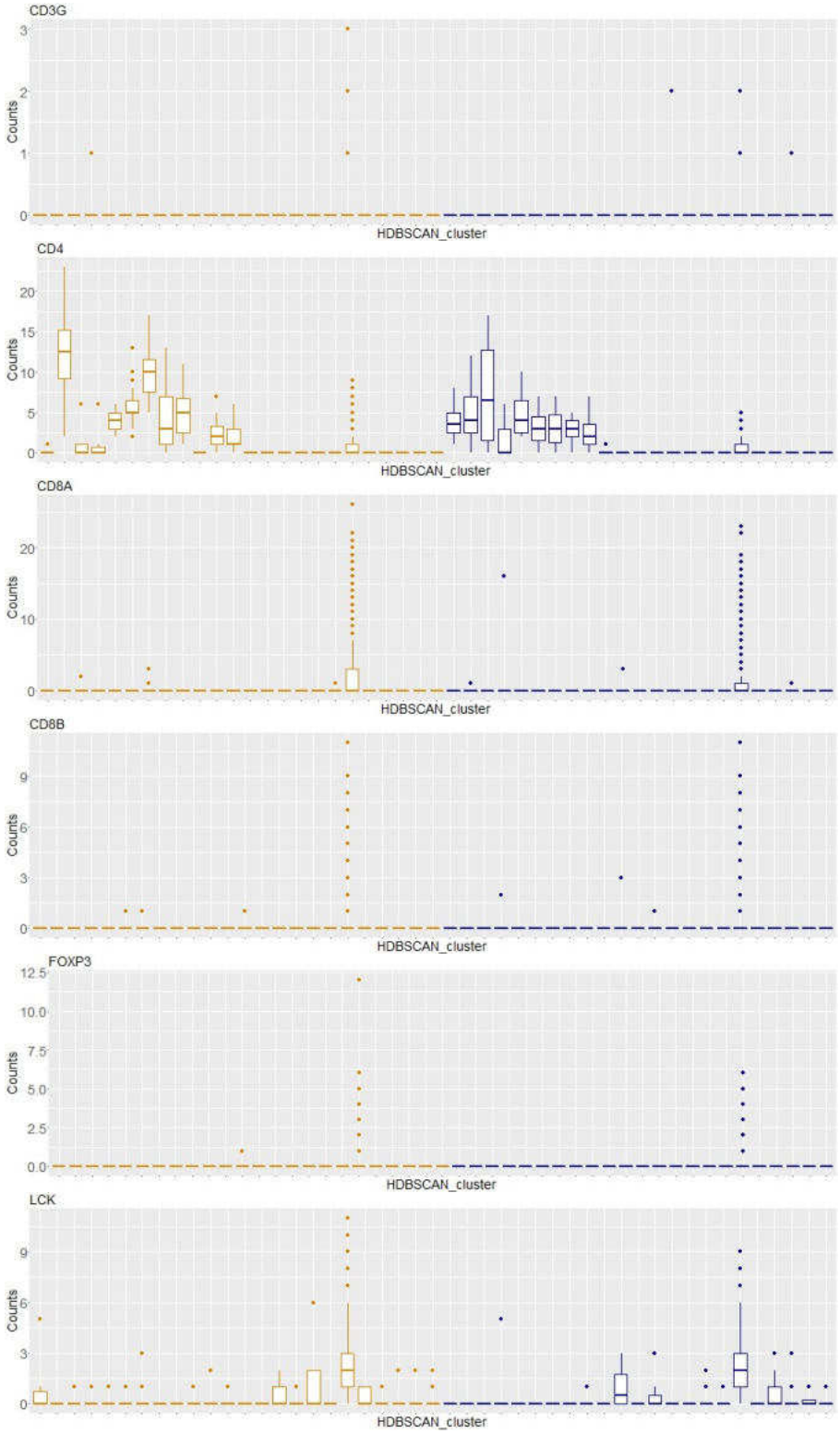
Monocytes

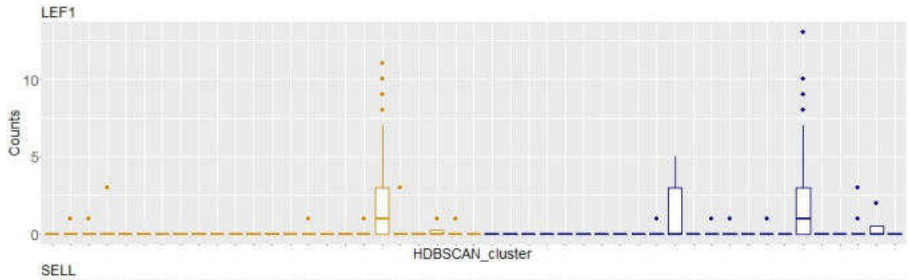


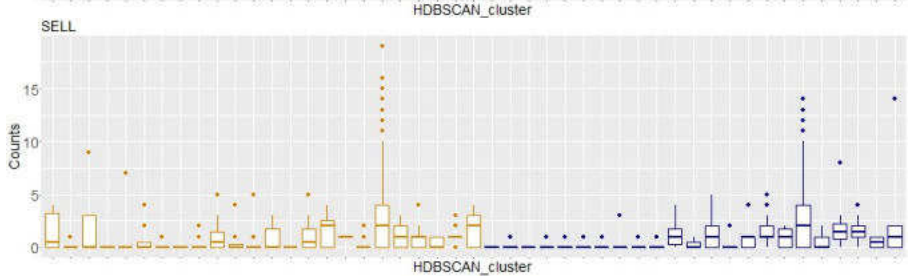
Dendritic cells



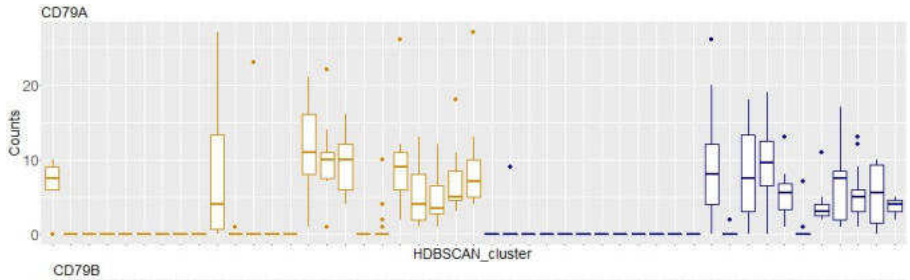
T cells

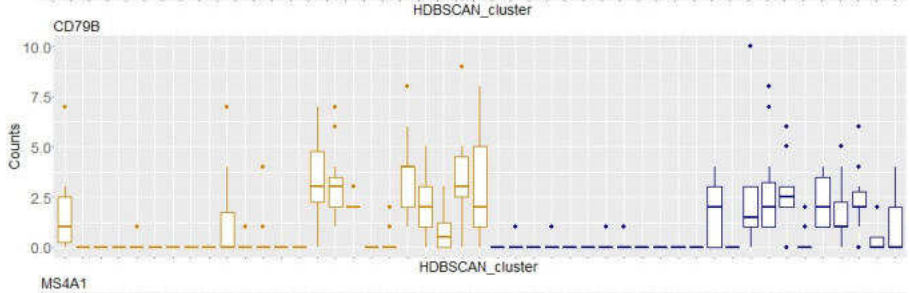

CD3G
Counts
HDBSCAN_cluster
CD4
Counts
HDBSCAN_cluster
CD8A
Counts
HDBSCAN_cluster
CD8B
Counts
HDBSCAN_cluster
FOXP3
Counts
HDBSCAN_cluster
LCK
Counts
HDBSCAN_cluster

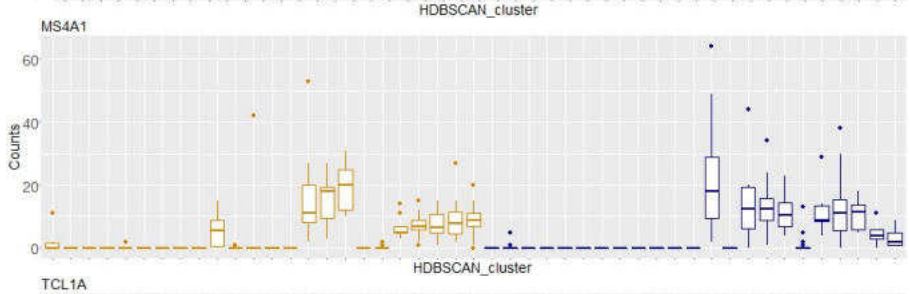

LEF1
Counts
HDBSCAN_cluster

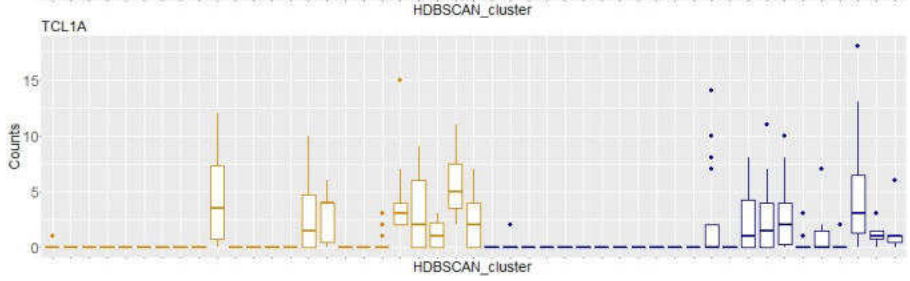

SELL
Counts
HDBSCAN_cluster

B cells

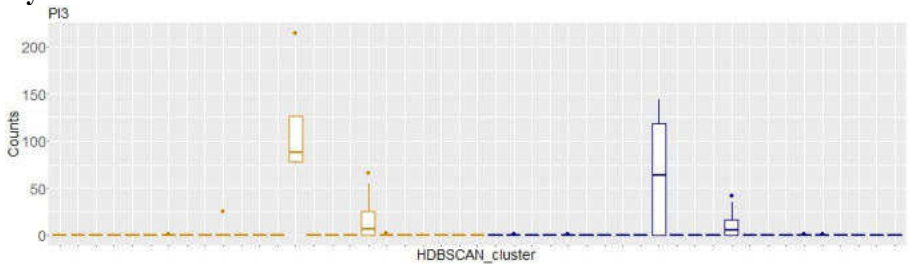

CD79A
Counts
HDBSCAN_cluster


CD79B
Counts
HDBSCAN_cluster

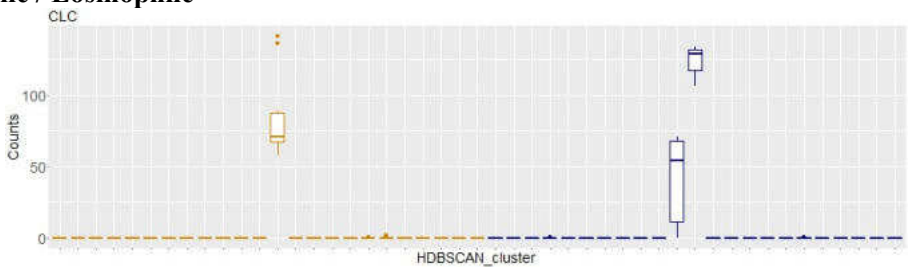

MS4A1
Counts
HDBSCAN_cluster


TCL1A
Counts
HDBSCAN_cluster

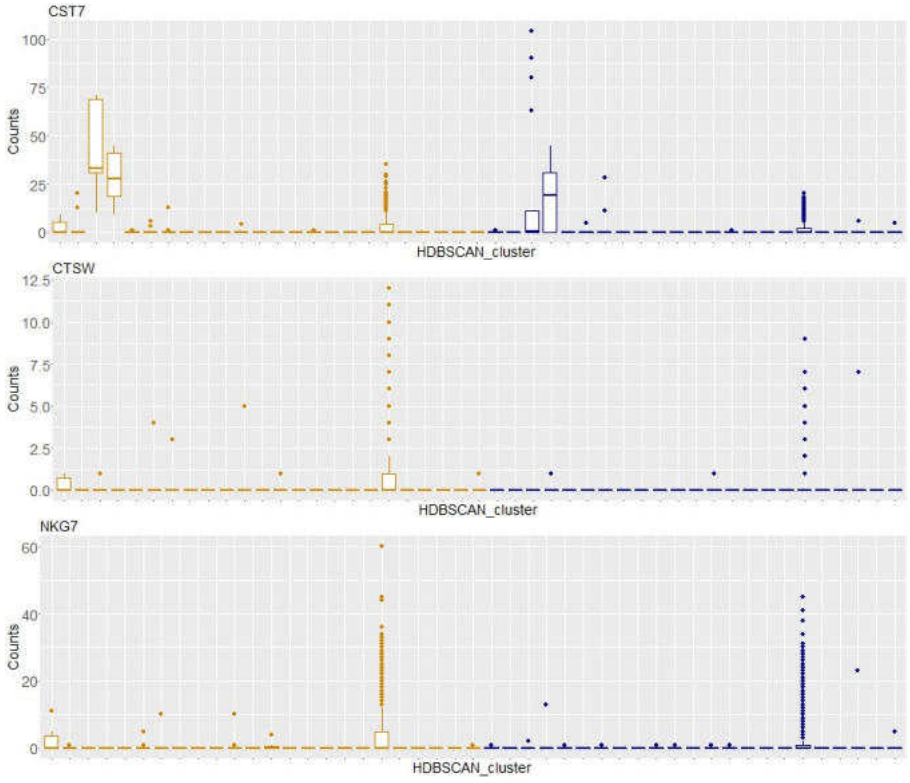
Granulocytes



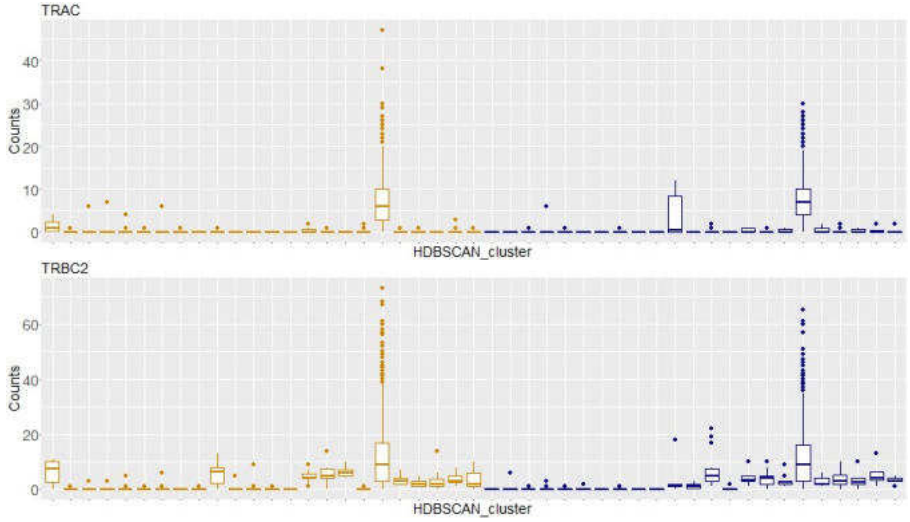
Basophile / Eosinophile



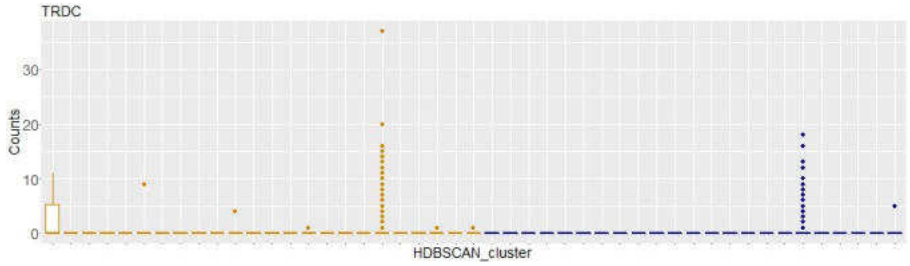
NK cells



Alpha / Beta T cells

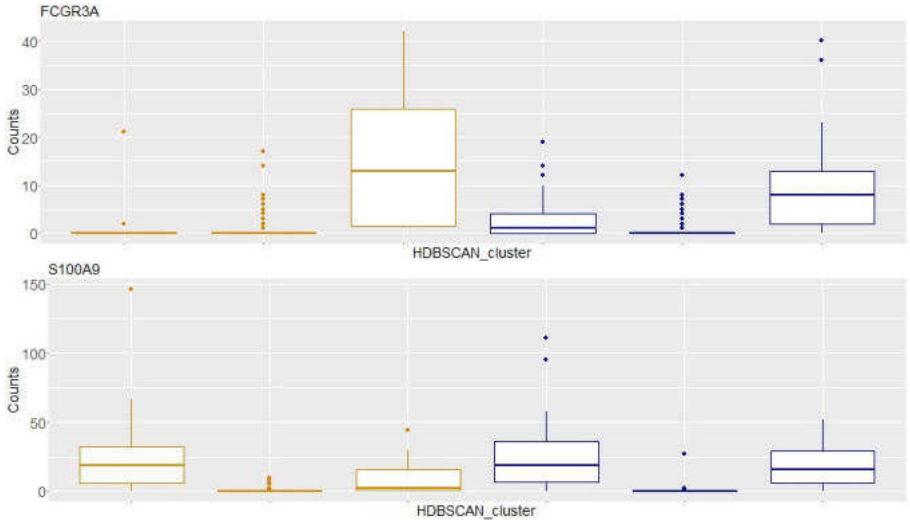


Gamma / Delta T cells

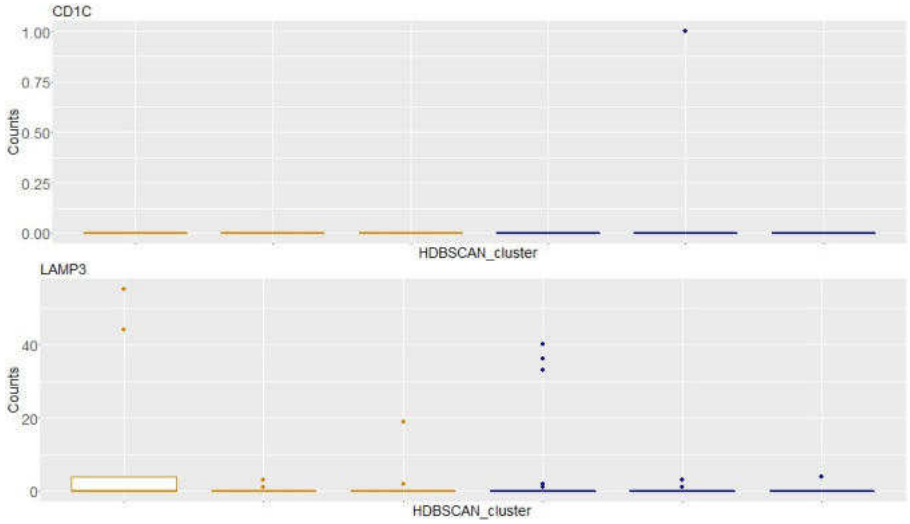


Set B test structure data

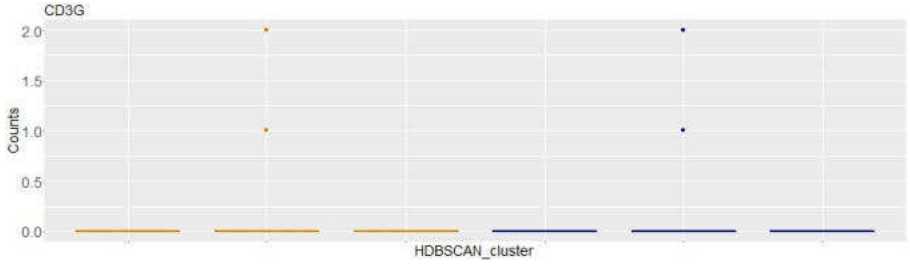
Monocytes

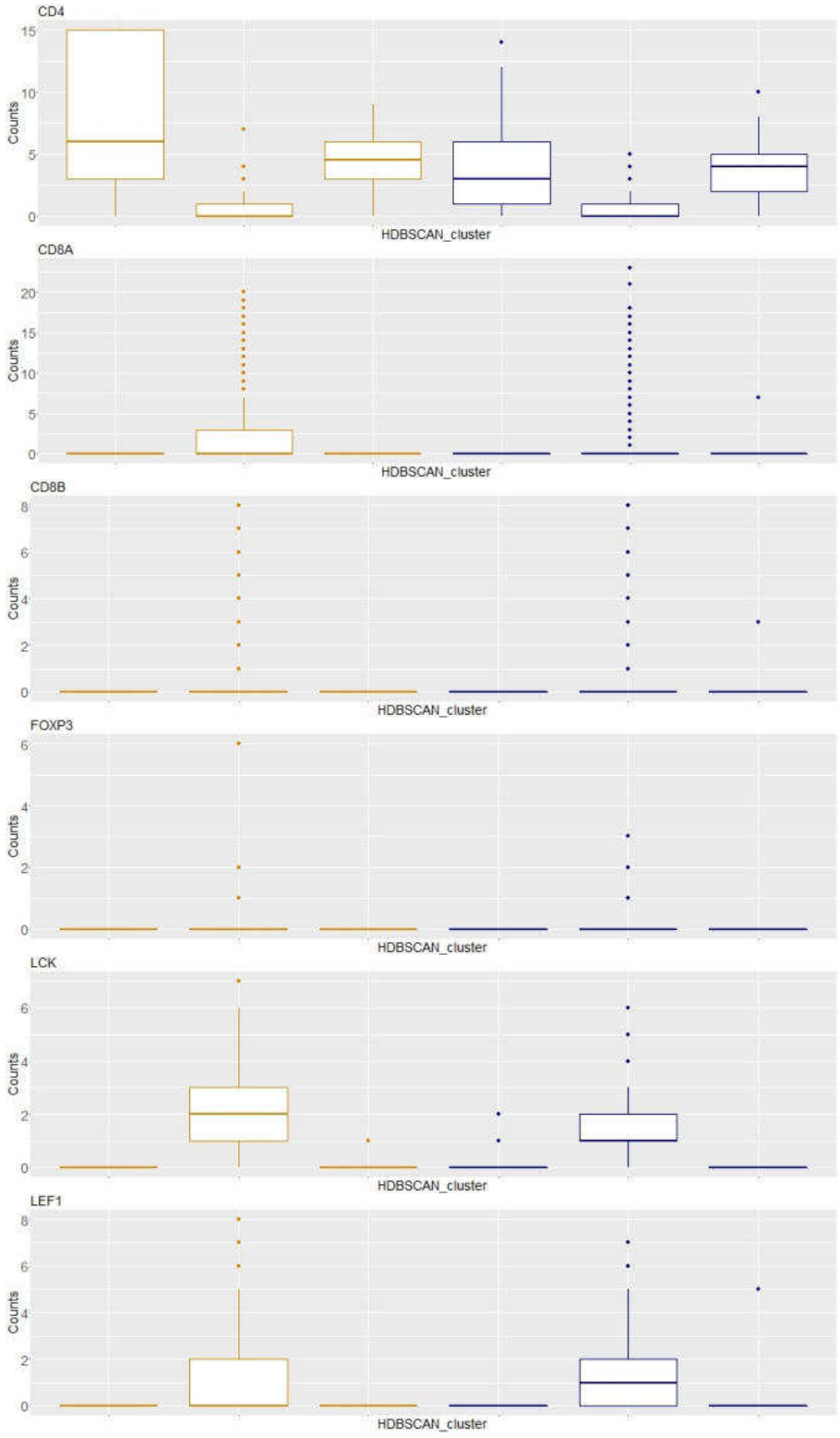


Dendritic cells

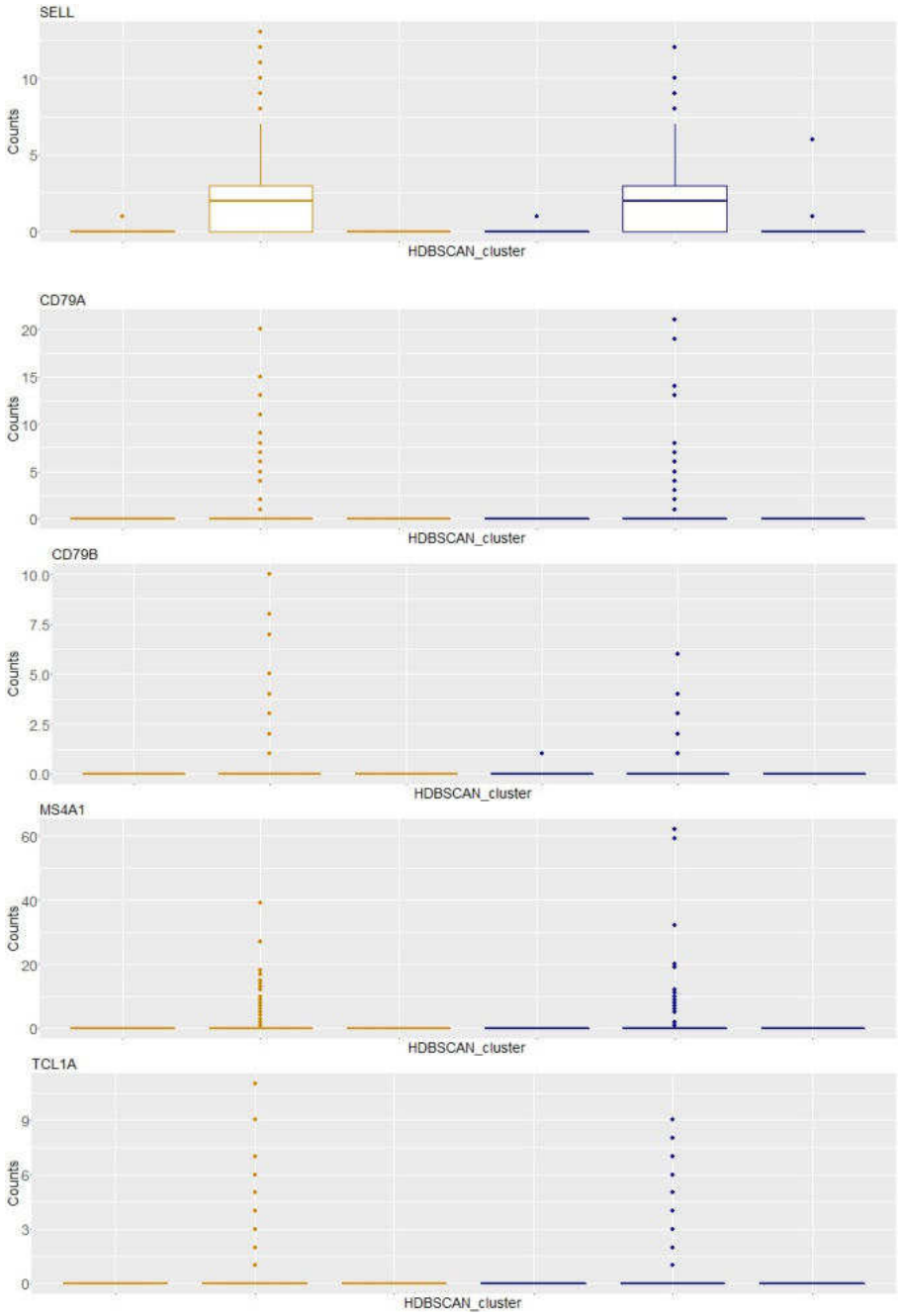


T cells

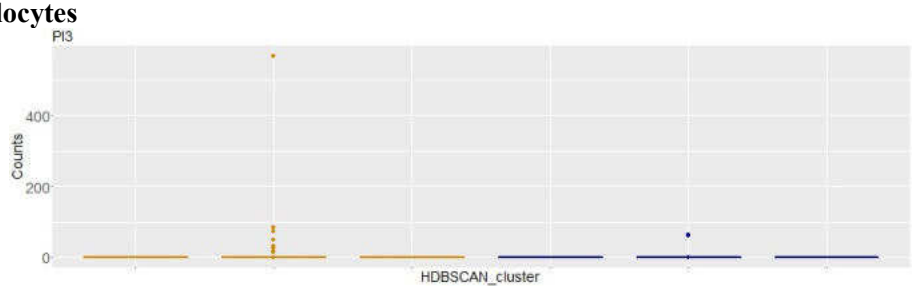



CD4
Counts
HDBSCAN_cluster
CD8A
Counts
HDBSCAN_cluster
CD8B
Counts
HDBSCAN_cluster
FOXP3
Counts
HDBSCAN_cluster
LCK
Counts
HDBSCAN_cluster
LEF1
Counts
HDBSCAN_cluster

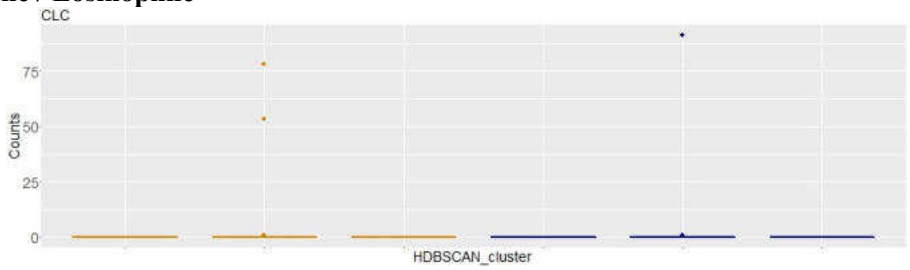
B cells



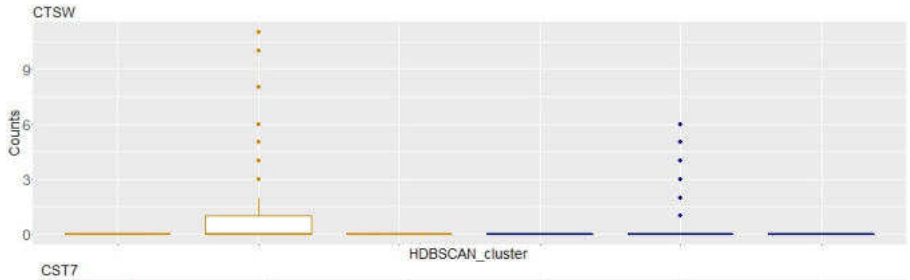
Granulocytes

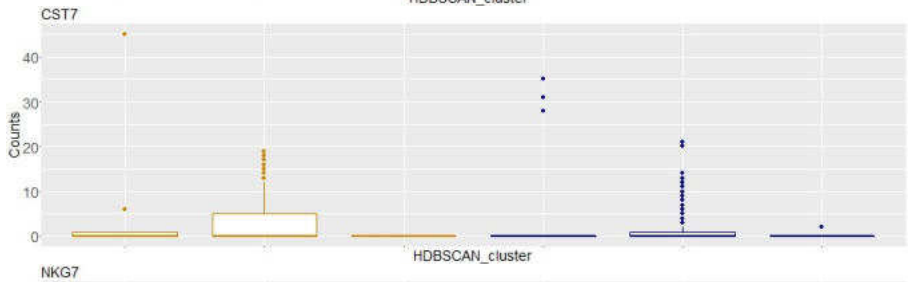


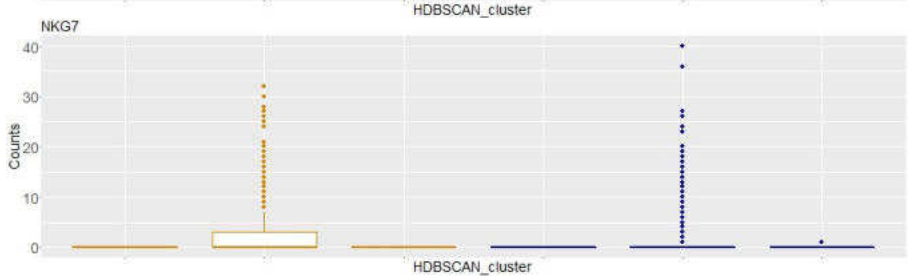
Basophile / Eosinophile



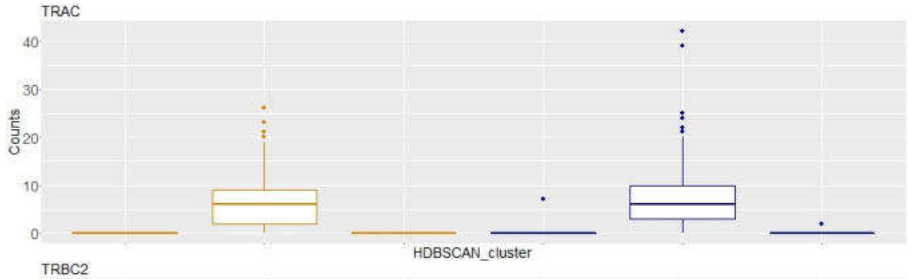
NK cells

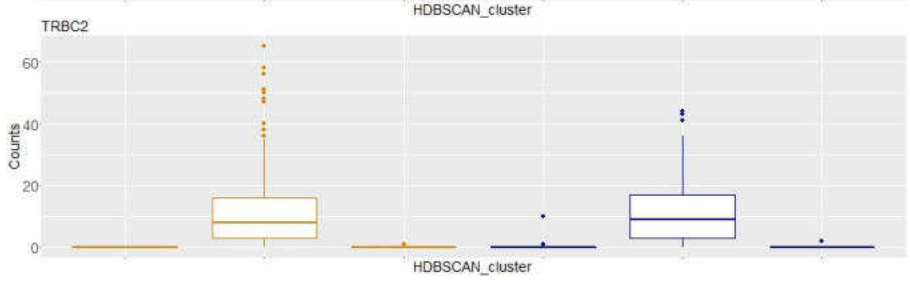




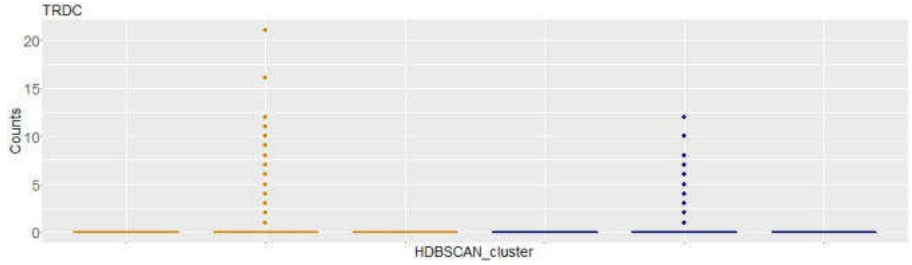


Alpha / Beta T cells





Gamma / Delta T cells

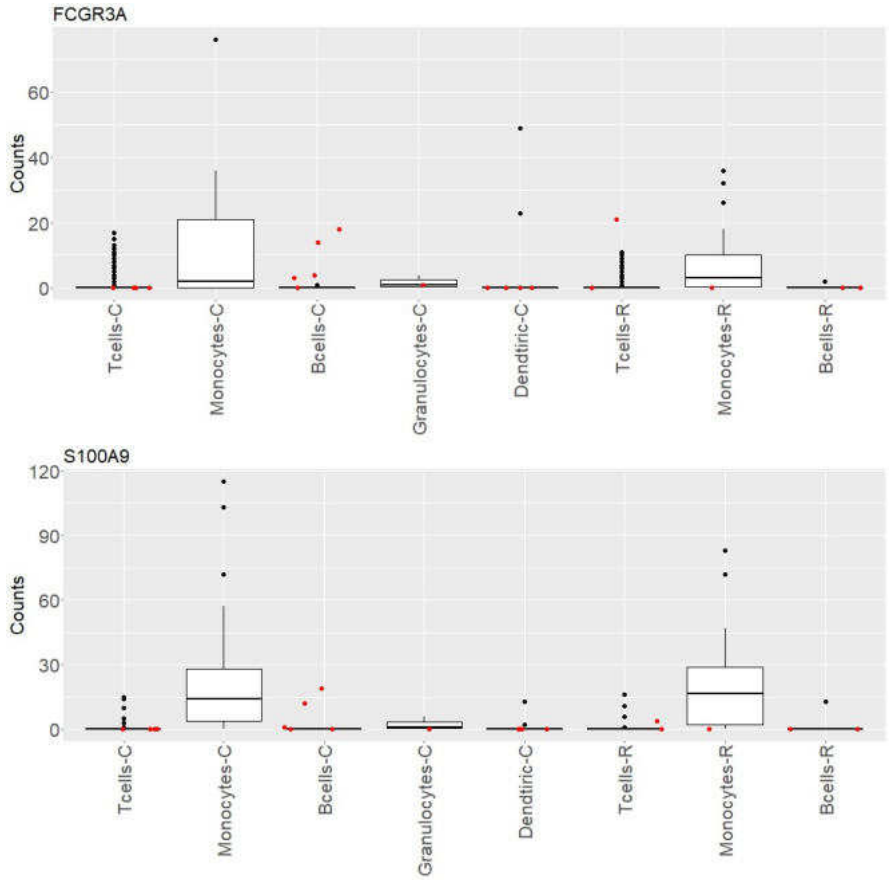


Recognition of suspicious cells affiliation

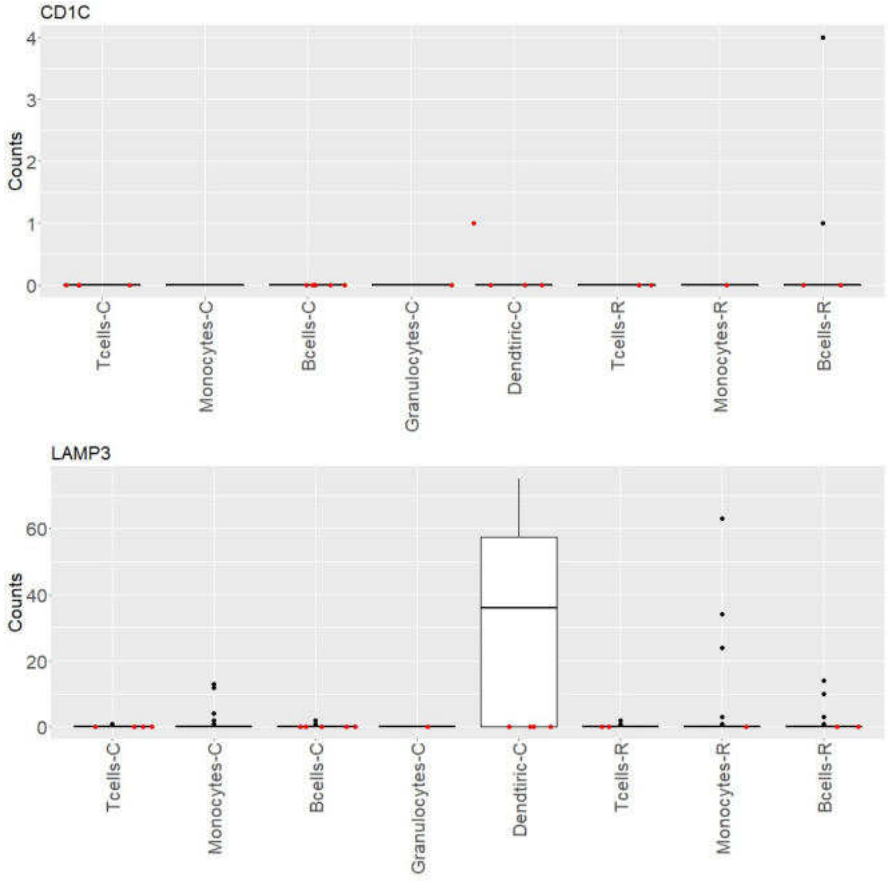
Analysis of suspicious cells required generating a series of count distribution boxplots for confirmed cell subpopulations. Suspicious cells are drawn separately and marked as red dots. Based on the figures below, it was decided to reclassify these cells to the other corresponding cell subpopulations. If within a particular cell there was noticed an increase in counts for a marker gene of a cell subpopulation other than the one to which it was assigned, that cell was removed from the current subpopulation and transferred to the corresponding subpopulation for which a significant increase in counts was observed. This section is divided into subsections depending on the analyzed data set. In connection with the problem, analysis required the model structures of the set A and set B data. These subsections were described with the name of the subpopulation for which the marker gene was analyzed.

Set A model structure data

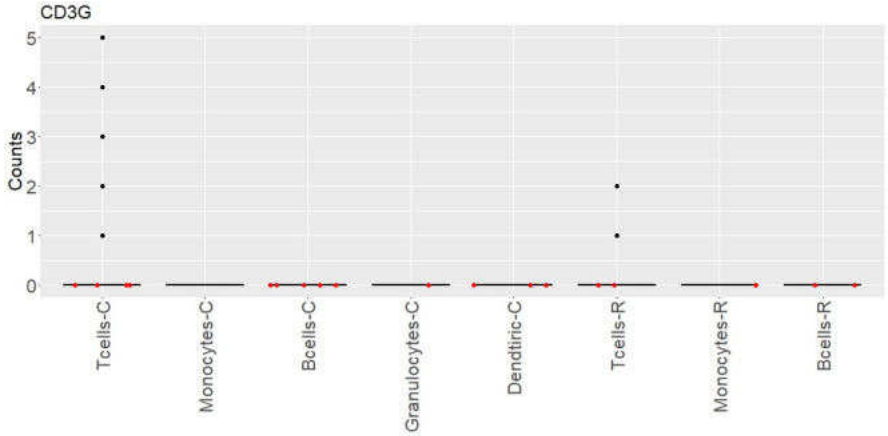
Monocytes

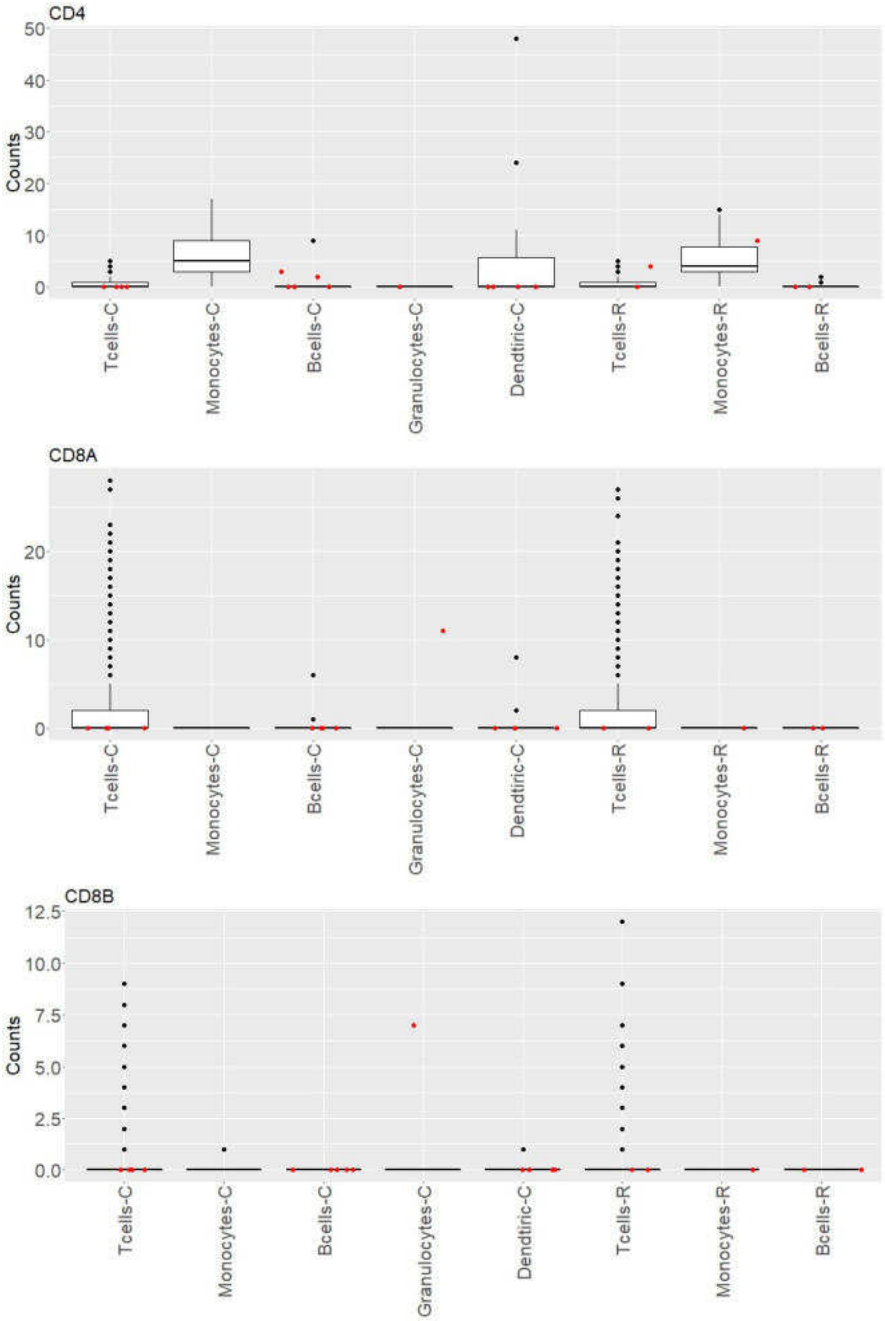


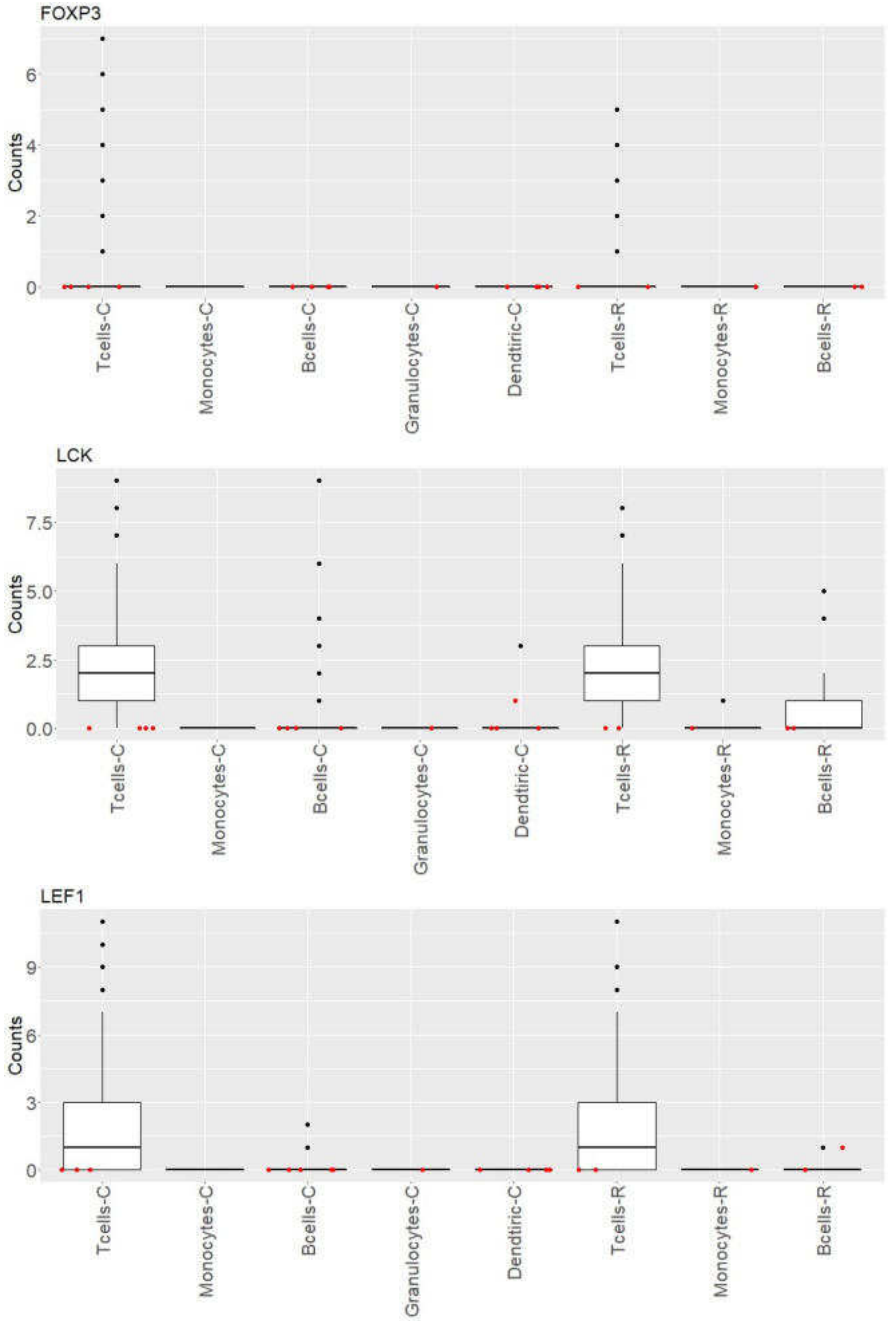
Dendritic cells

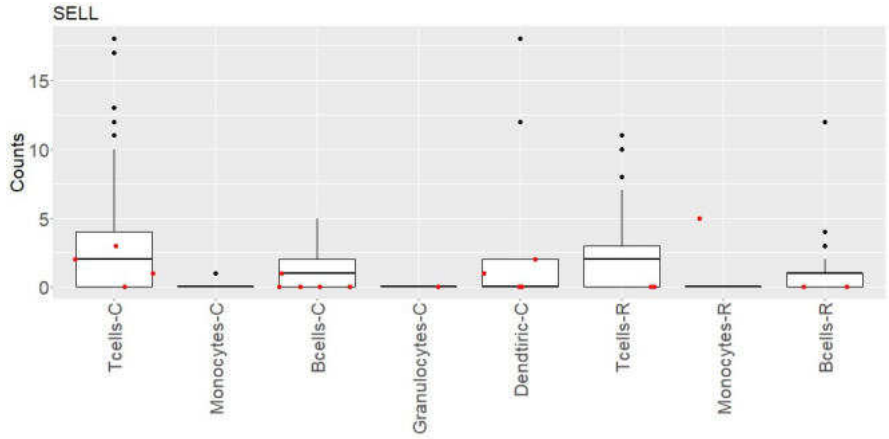


T cells

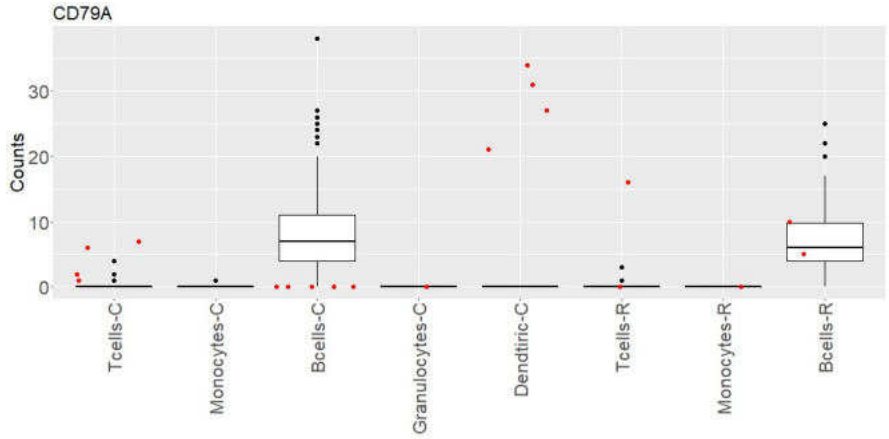


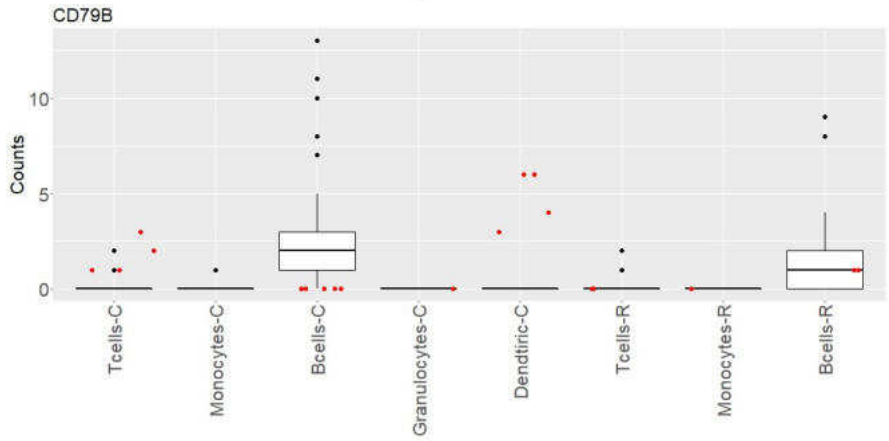

CD4
Counts
50
40
30
20
10
0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Tcells-R
Monocytes-R
Bcells-R
CD8A
Counts
20
10
0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Tcells-R
Monocytes-R
Bcells-R
CD8B
Counts
12.5
10.0
7.5
5.0
2.5
0.0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Tcells-R
Monocytes-R
Bcells-R

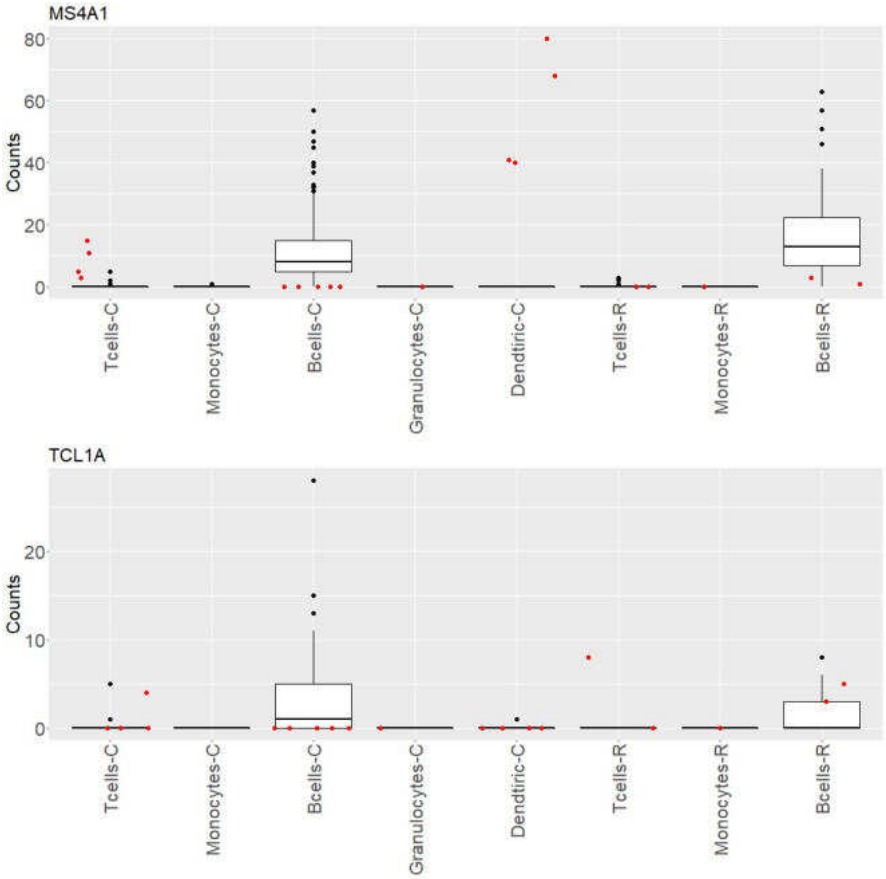

FOXP3
Counts
0
2
4
6
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Tcells-R
Monocytes-R
Bcells-R
LCK
Counts
0.0
2.5
5.0
7.5
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Tcells-R
Monocytes-R
Bcells-R
LEF1
Counts
0
3
6
9
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Tcells-R
Monocytes-R
Bcells-R


SELL
Counts
15
10
5
0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Tcells-R
Monocytes-R
Bcells-R

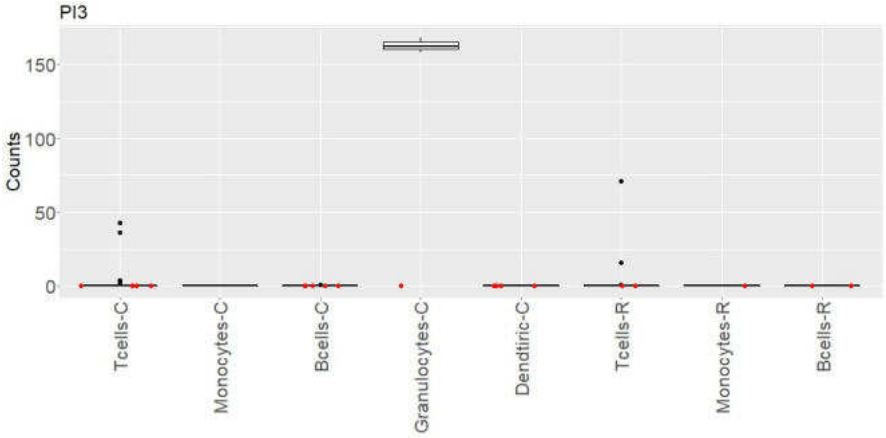
B cells


CD79A
Counts
30
20
10
0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Tcells-R
Monocytes-R
Bcells-R


CD79B
Counts
10
5
0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Tcells-R
Monocytes-R
Bcells-R

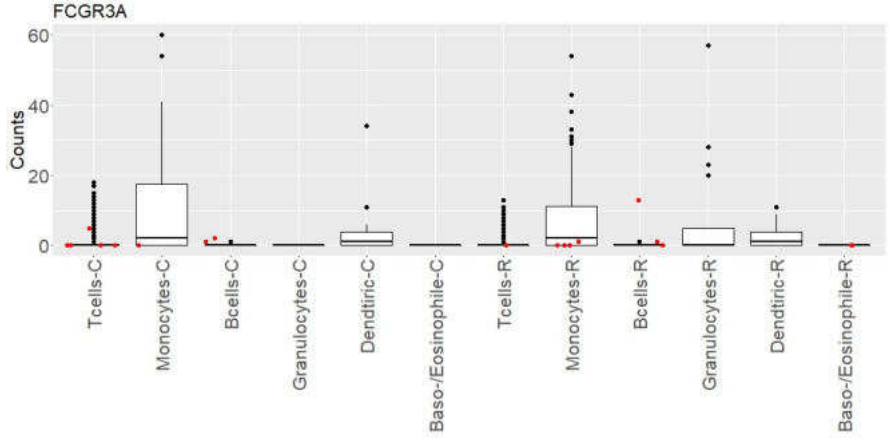


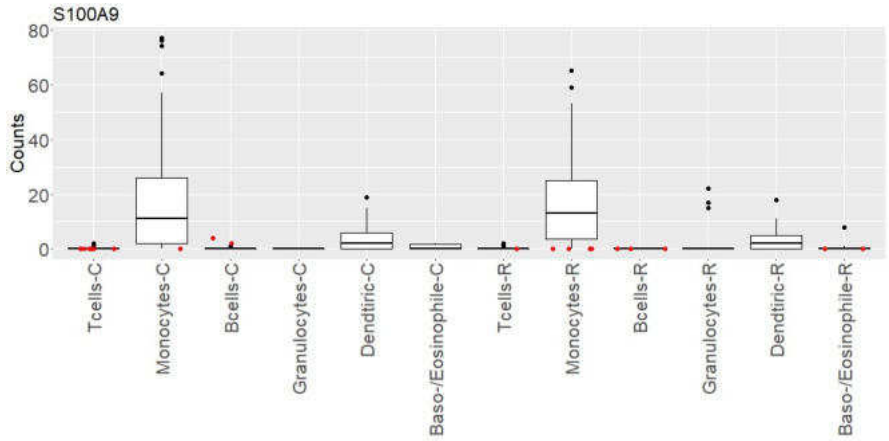
Granulocytes



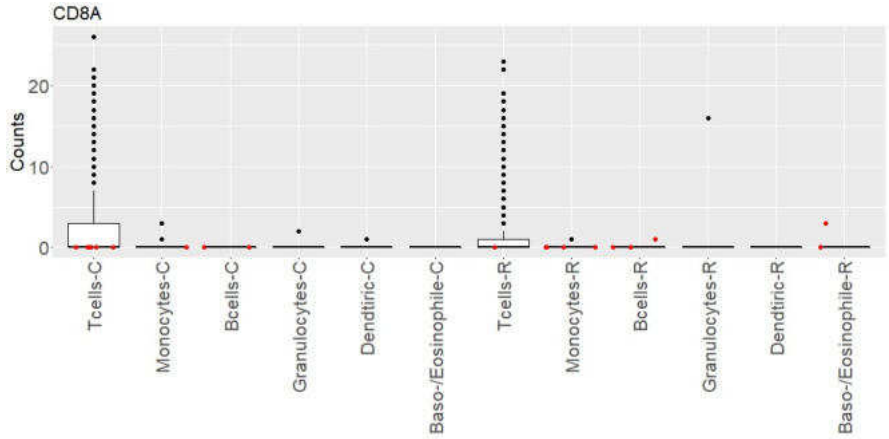
Set B model structure data

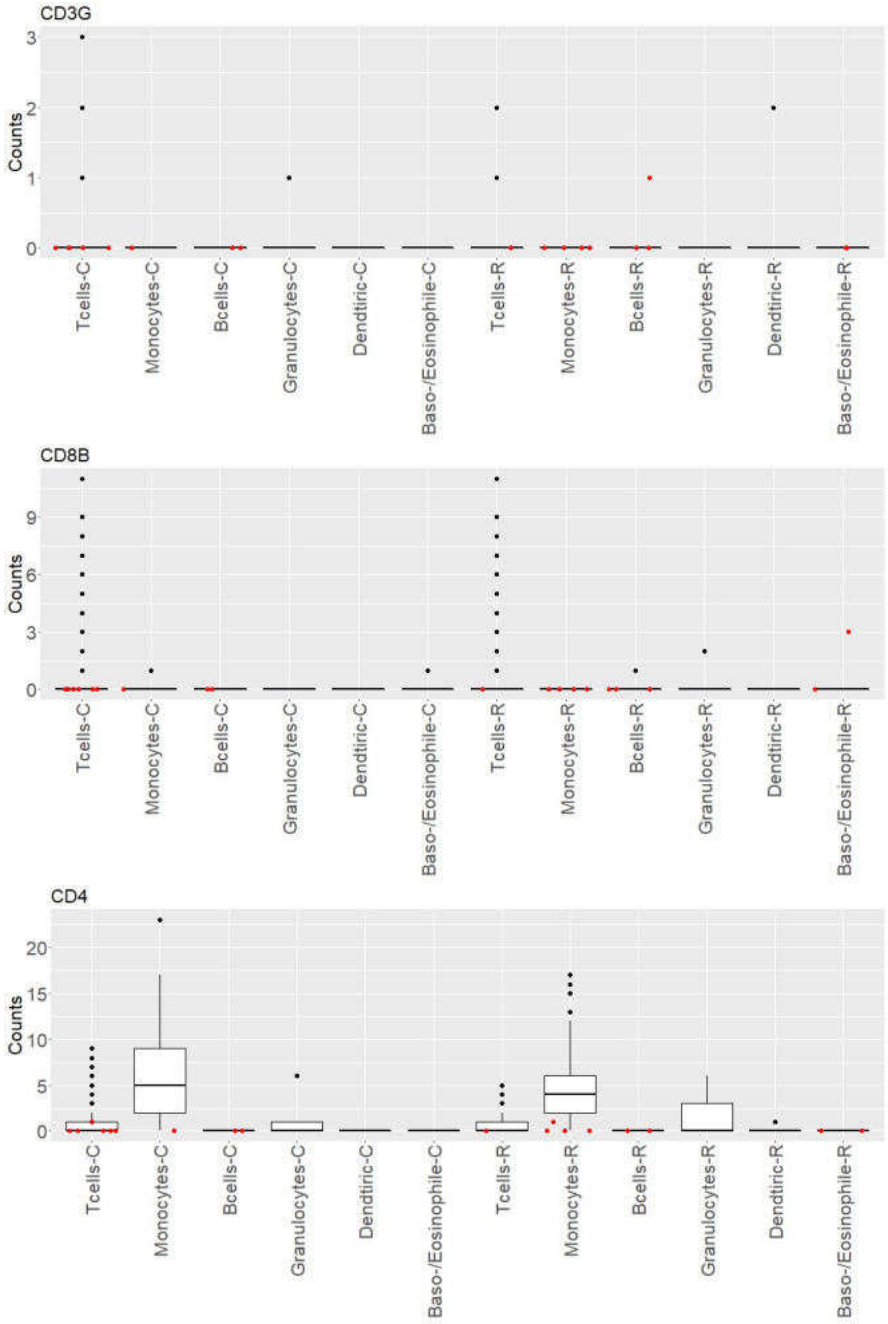
Monocytes

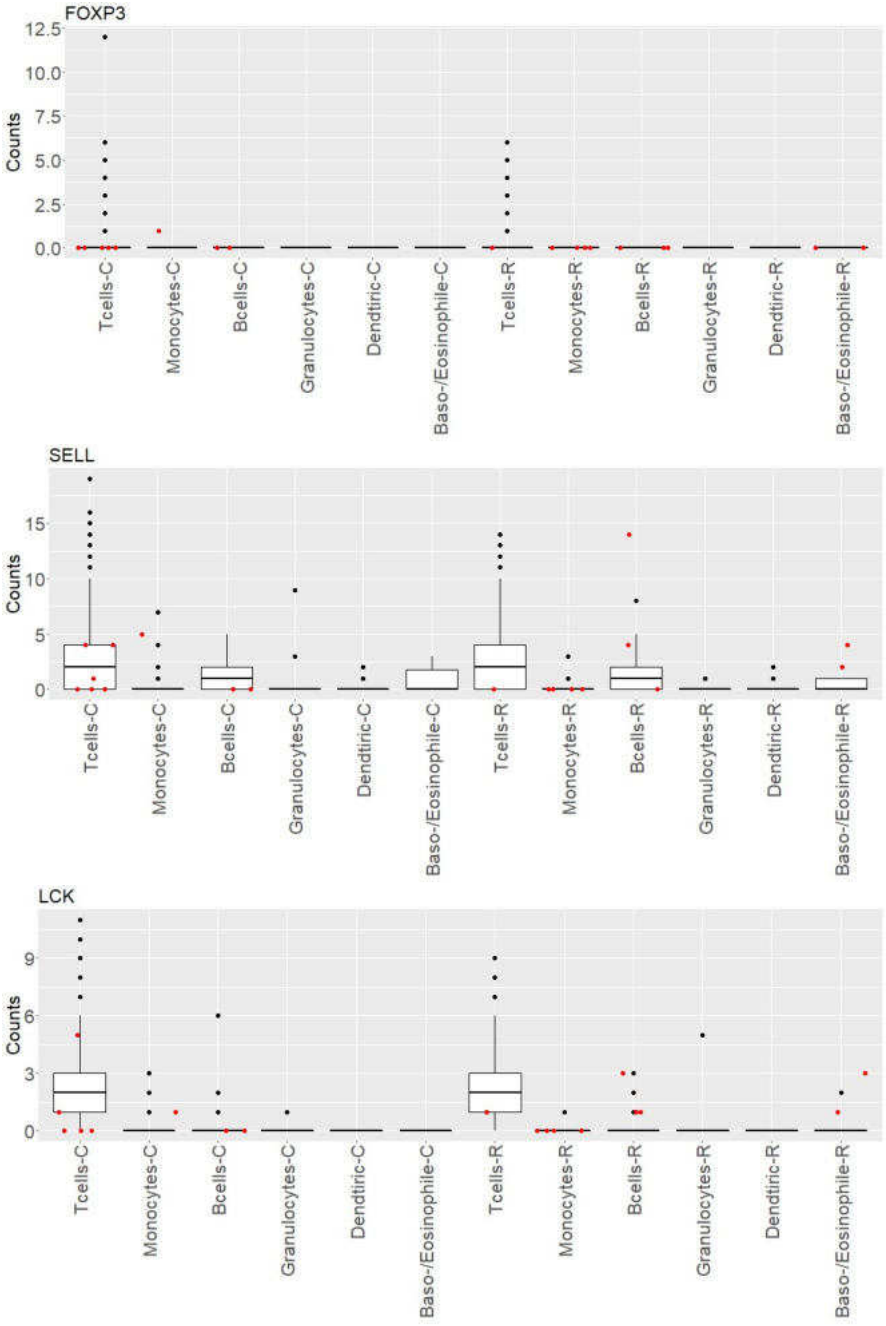


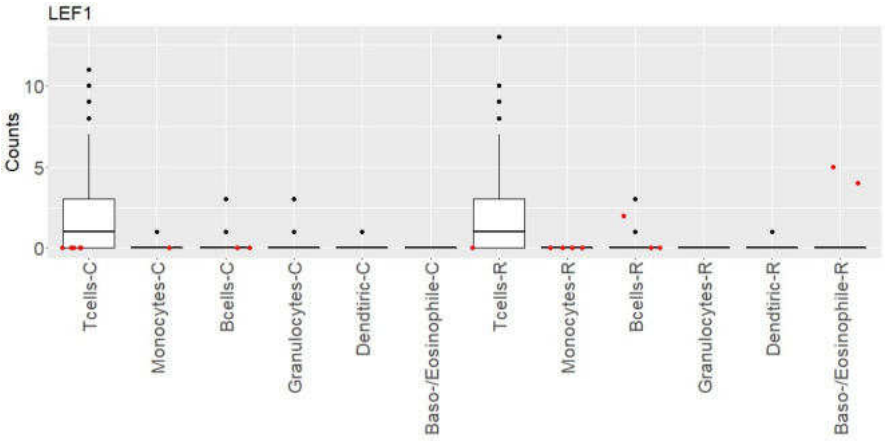


T cells

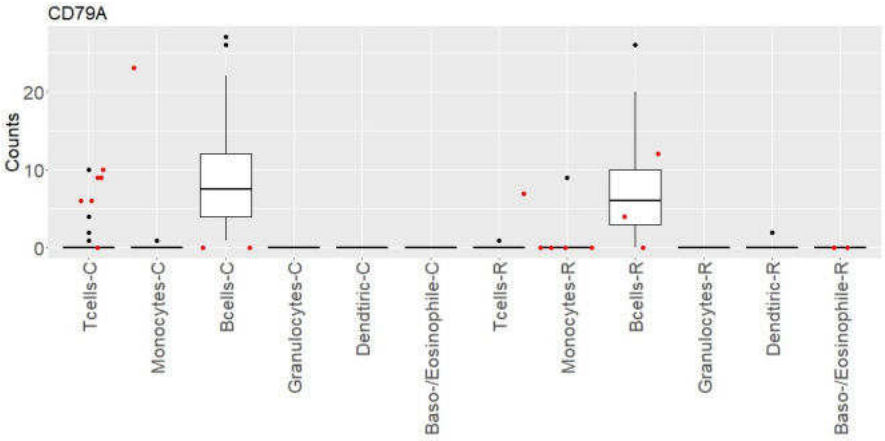


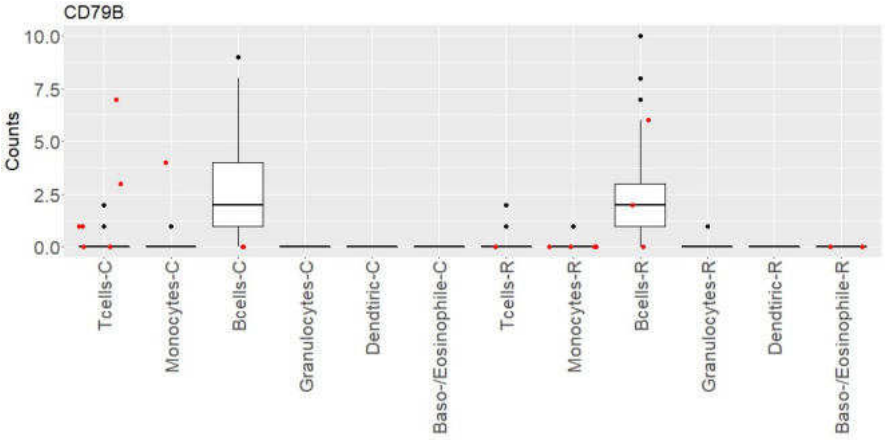

CD3G
Counts
0
1
2
3
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Baso-/Eosinophile-C
Tcells-R
Monocytes-R
Bcells-R
Granulocytes-R
Dendtiric-R
Baso-/Eosinophile-R
CD8B
Counts
0
3
6
9
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Baso-/Eosinophile-C
Tcells-R
Monocytes-R
Bcells-R
Granulocytes-R
Dendtiric-R
Baso-/Eosinophile-R
CD4
Counts
0
5
10
15
20
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Baso-/Eosinophile-C
Tcells-R
Monocytes-R
Bcells-R
Granulocytes-R
Dendtiric-R
Baso-/Eosinophile-R

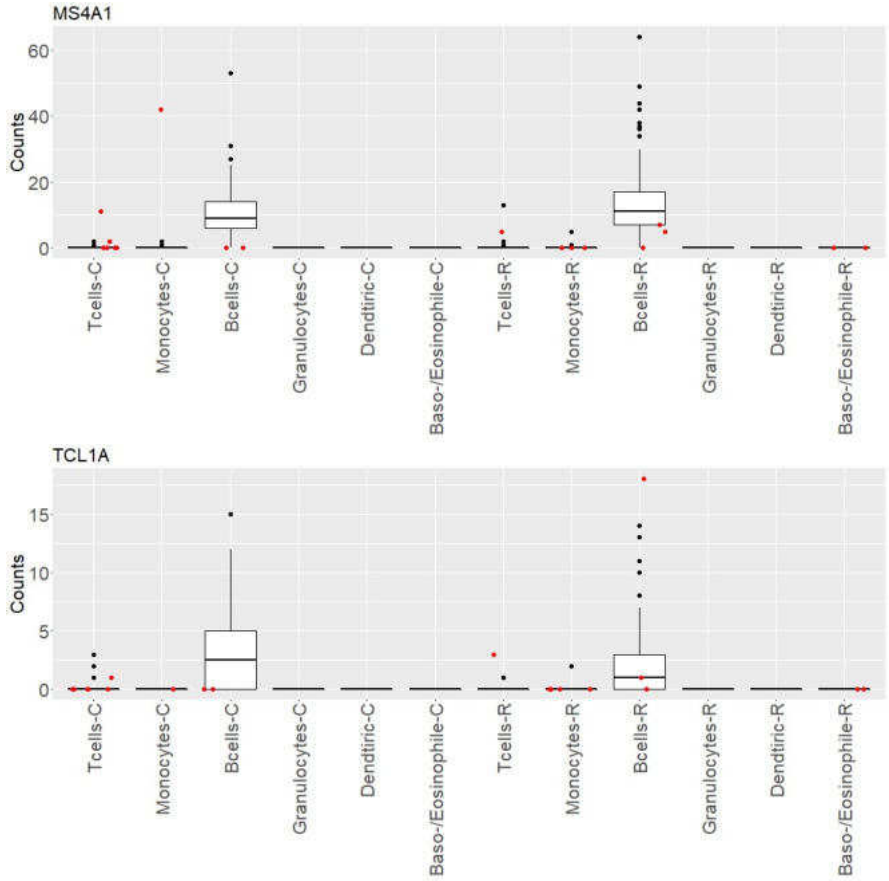

FOXP3
Counts
12.5
10.0
7.5
5.0
2.5
0.0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Baso-/Eosinophile-C
Tcells-R
Monocytes-R
Bcells-R
Granulocytes-R
Dendtiric-R
Baso-/Eosinophile-R
SELL
Counts
15
10
5
0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Baso-/Eosinophile-C
Tcells-R
Monocytes-R
Bcells-R
Granulocytes-R
Dendtiric-R
Baso-/Eosinophile-R
LCK
Counts
9
6
3
0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Baso-/Eosinophile-C
Tcells-R
Monocytes-R
Bcells-R
Granulocytes-R
Dendtiric-R
Baso-/Eosinophile-R


LEF1
Counts
10
5
0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Baso-/Eosinophile-C
Tcells-R
Monocytes-R
Bcells-R
Granulocytes-R
Dendtiric-R
Baso-/Eosinophile-R

B cells

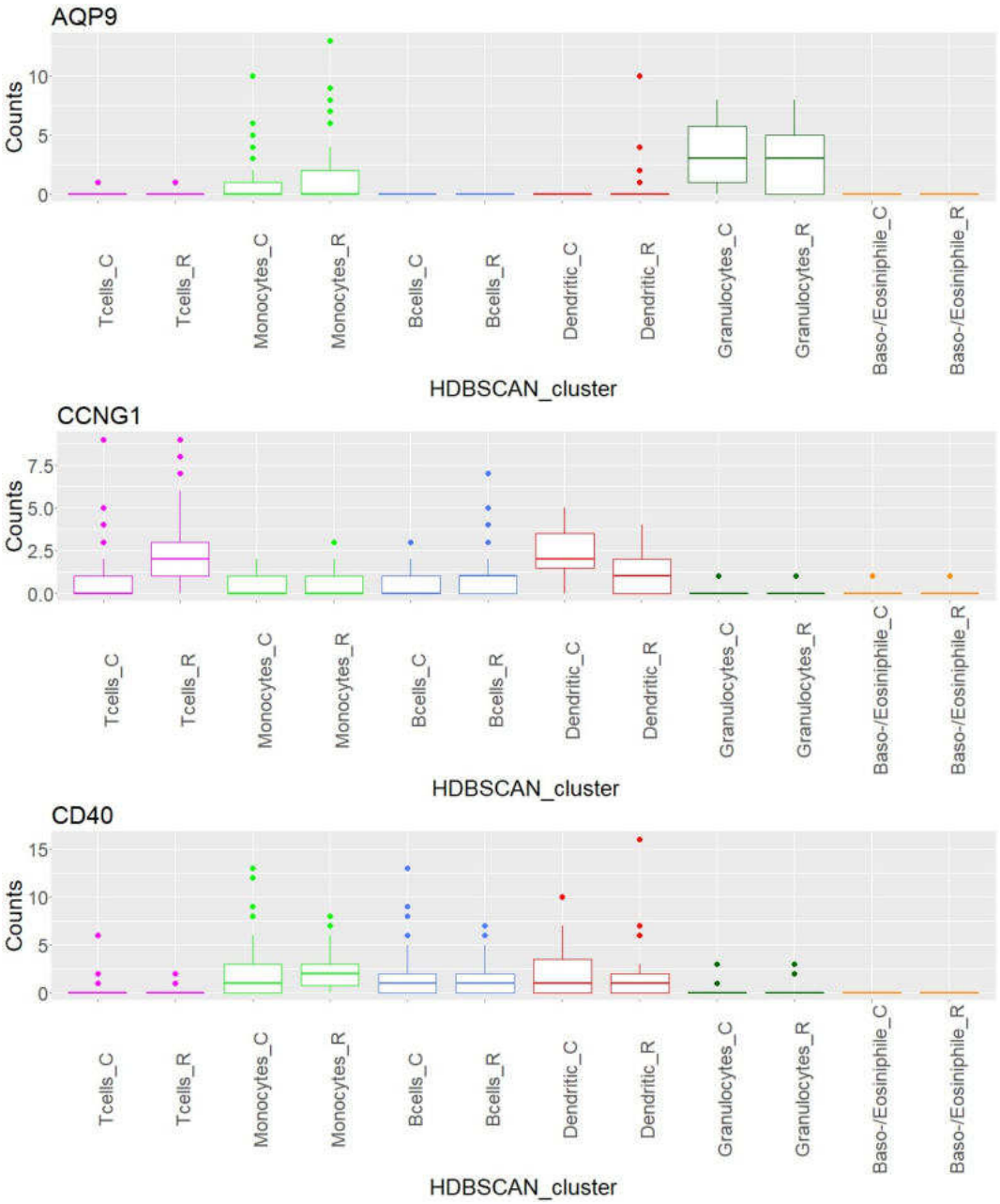

CD79A
Counts
20
10
0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Baso-/Eosinophile-C
Tcells-R
Monocytes-R
Bcells-R
Granulocytes-R
Dendtiric-R
Baso-/Eosinophile-R

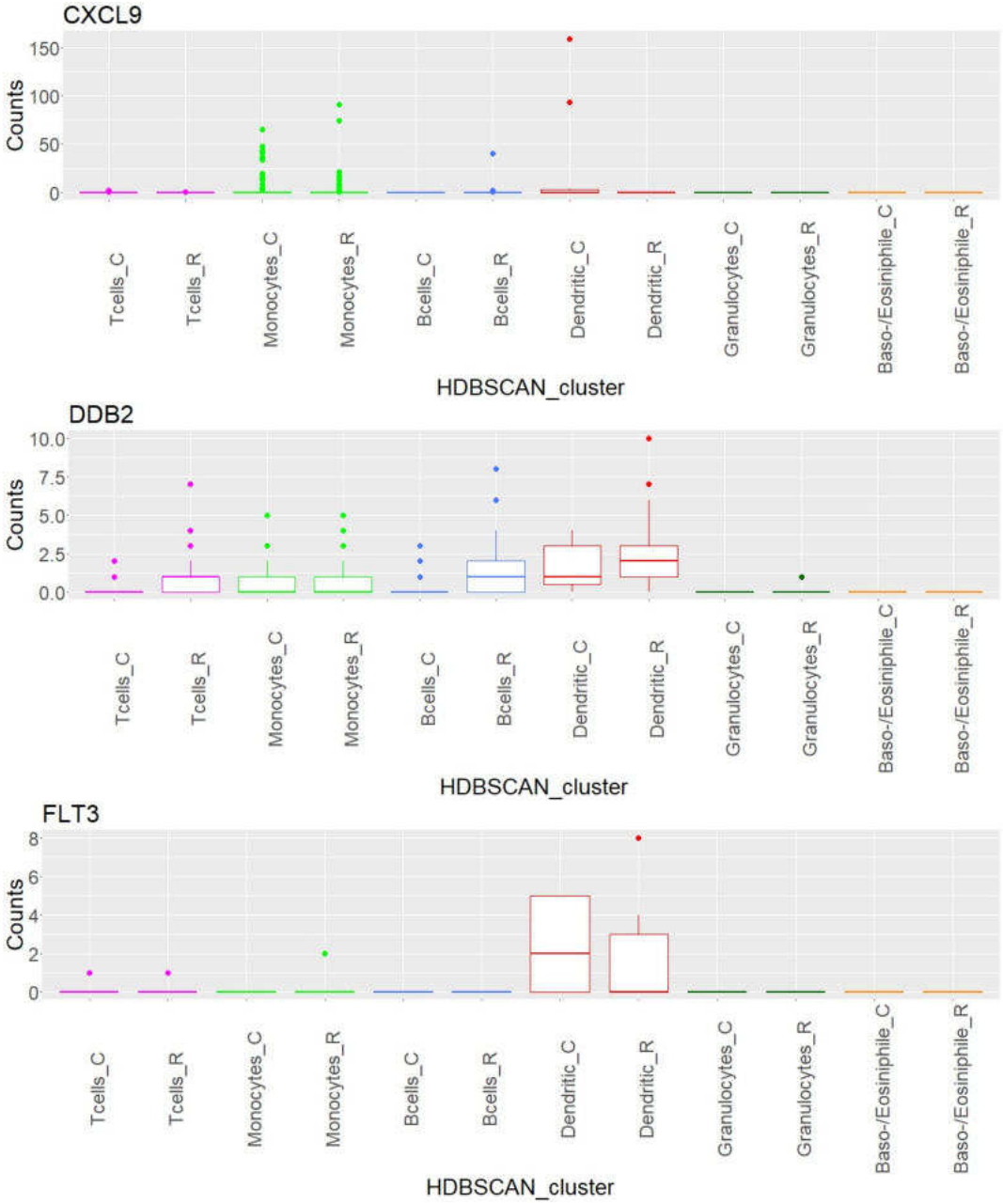

CD79B
Counts
10.0
7.5
5.0
2.5
0.0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Baso-/Eosinophile-C
Tcells-R
Monocytes-R
Bcells-R
Granulocytes-R
Dendtiric-R
Baso-/Eosinophile-R

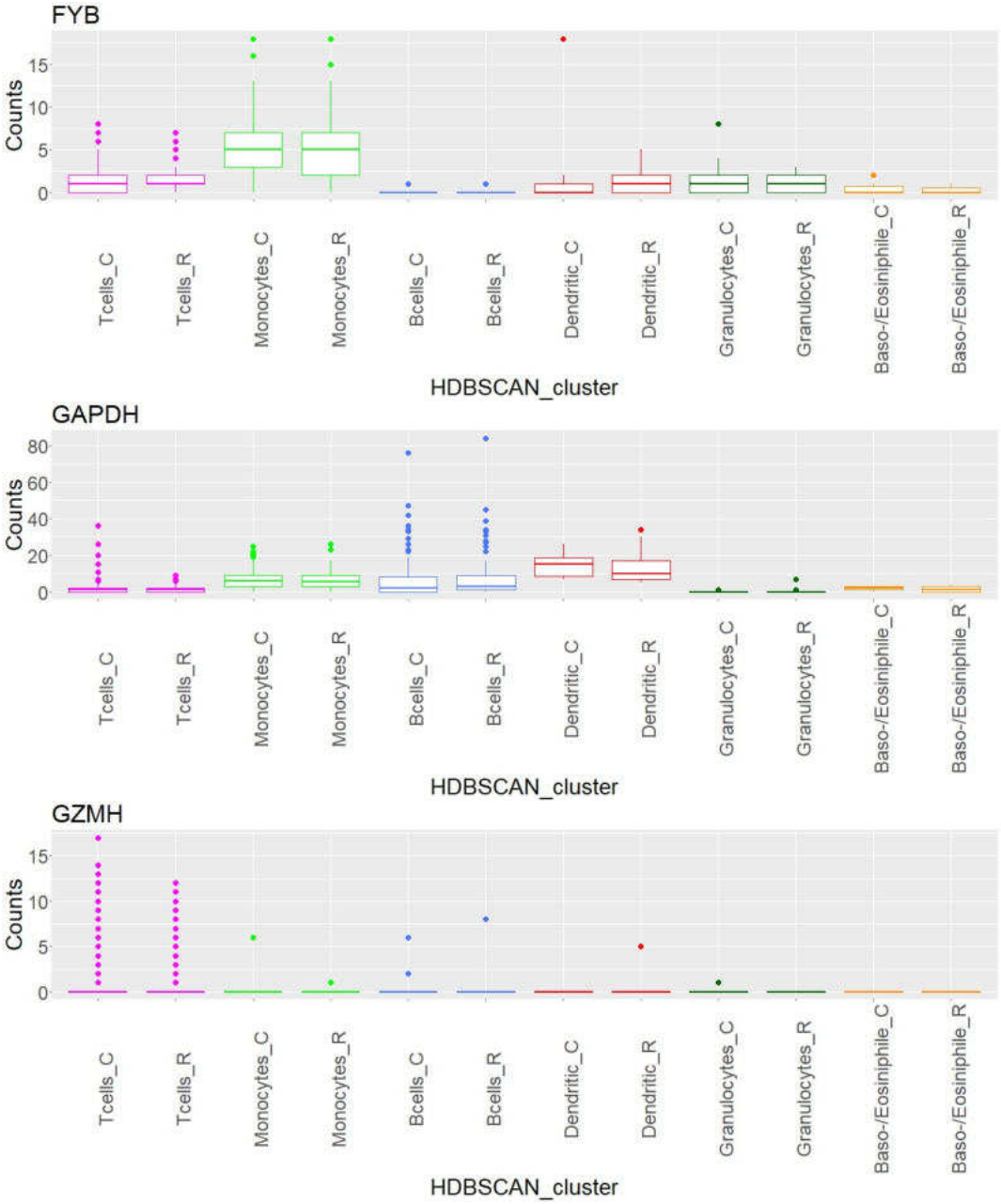

MS4A1
Counts
60
40
20
0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Baso-/Eosinophile-C
Tcells-R
Monocytes-R
Bcells-R
Granulocytes-R
Dendtiric-R
Baso-/Eosinophile-R
TCL1A
Counts
15
10
5
0
Tcells-C
Monocytes-C
Bcells-C
Granulocytes-C
Dendtiric-C
Baso-/Eosinophile-C
Tcells-R
Monocytes-R
Bcells-R
Granulocytes-R
Dendtiric-R
Baso-/Eosinophile-R

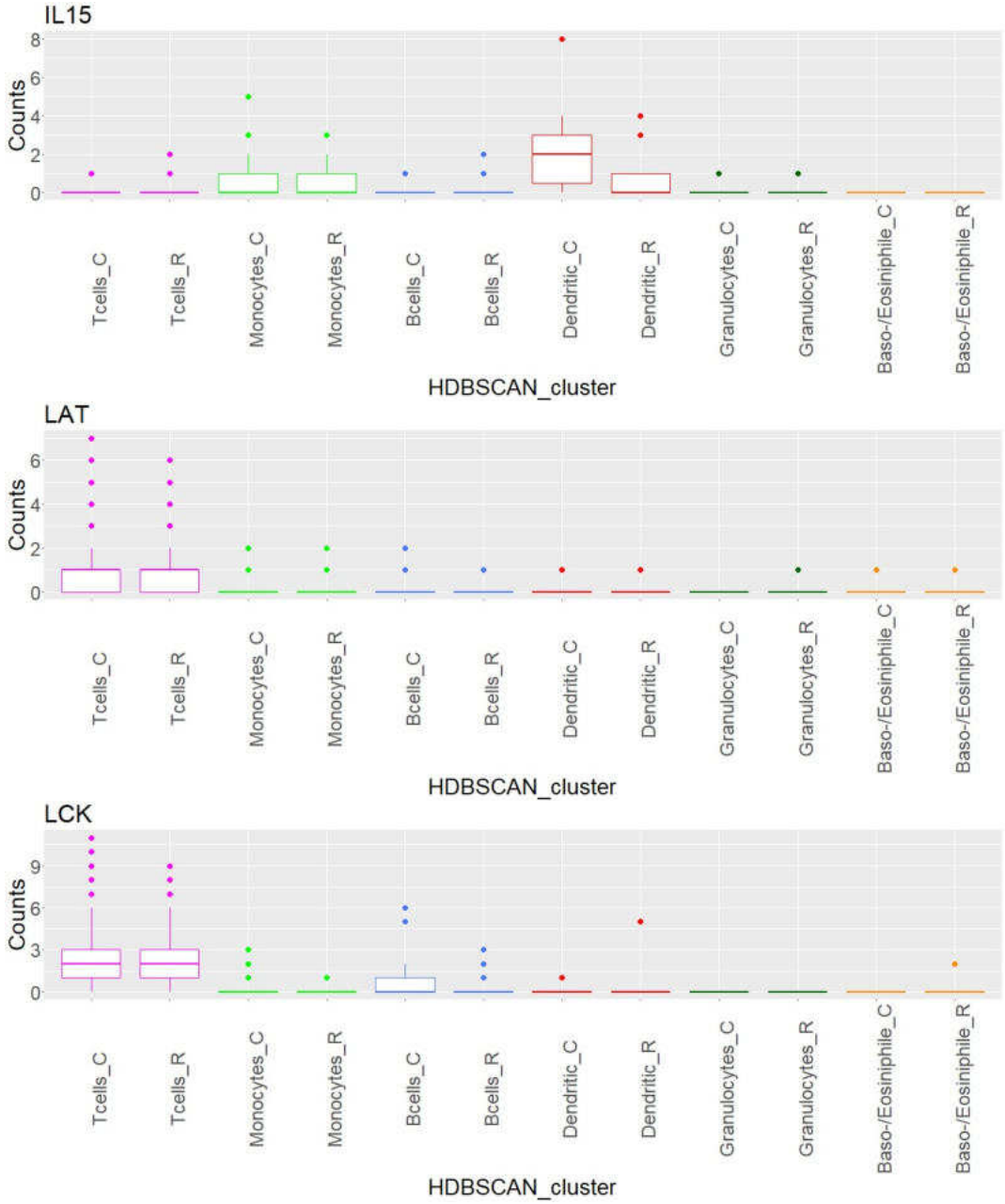
Distribution of counts for selected genes by cell subpopulations

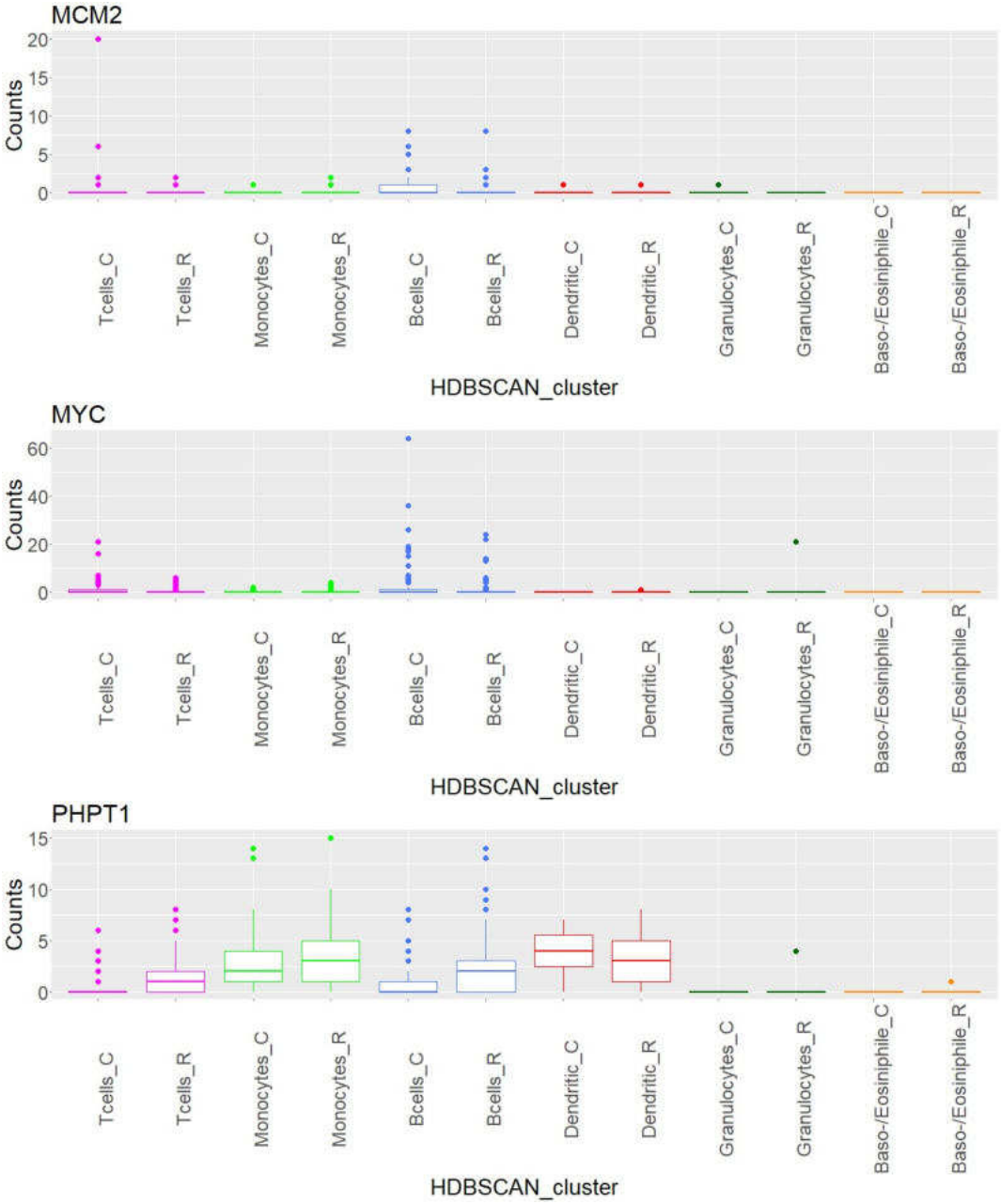
Additional analysis of count distribution for individual genes chosen in the feature selection process was intended to indicate the presence of genes associated with high heterogeneity of the analyzed dataset. Distribution boxplots were created based on the set B model structure. Individual cell subpopulations have been marked with a consistent color map. Control samples are marked with a C, and irradiated samples with an R in the name.

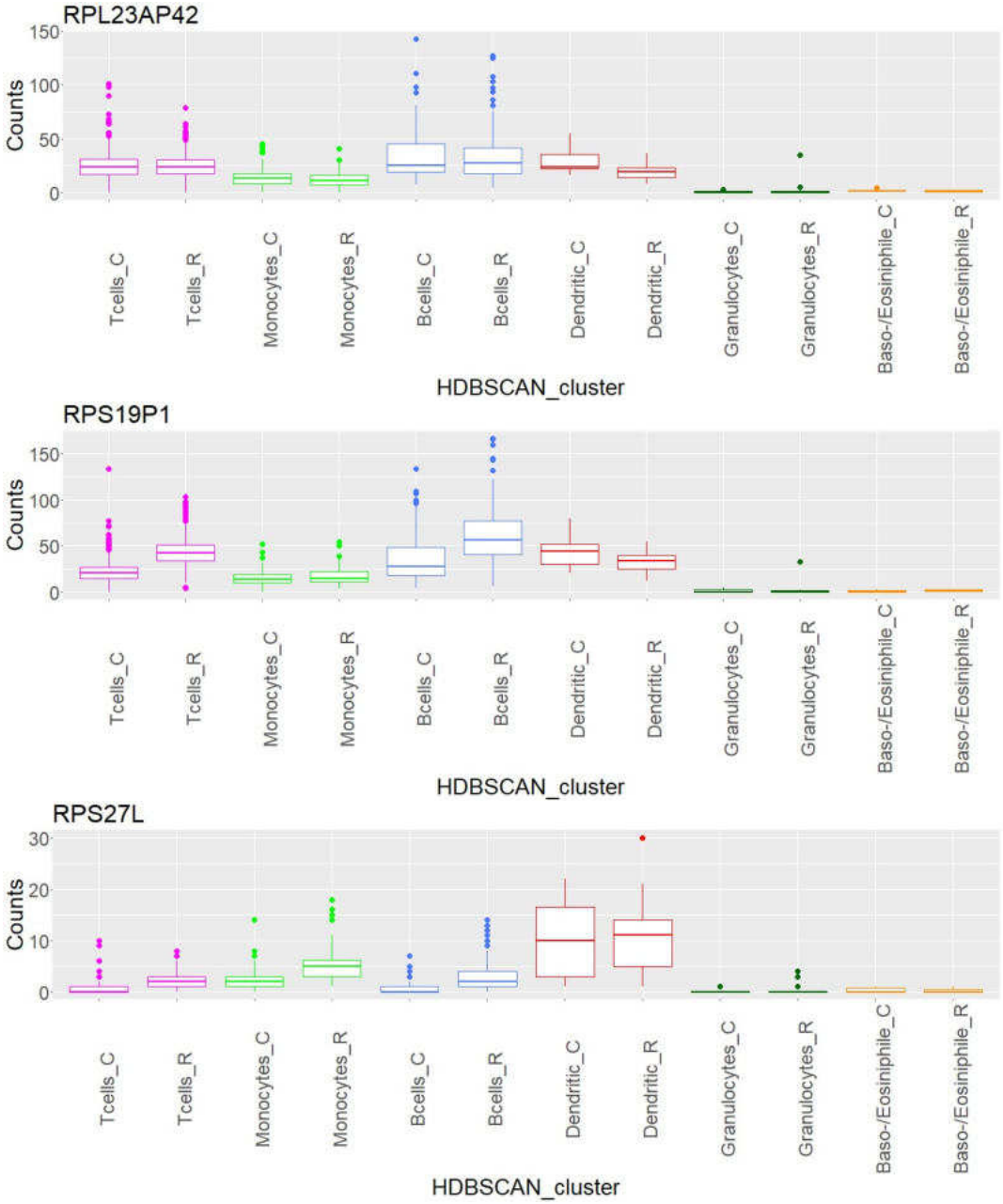


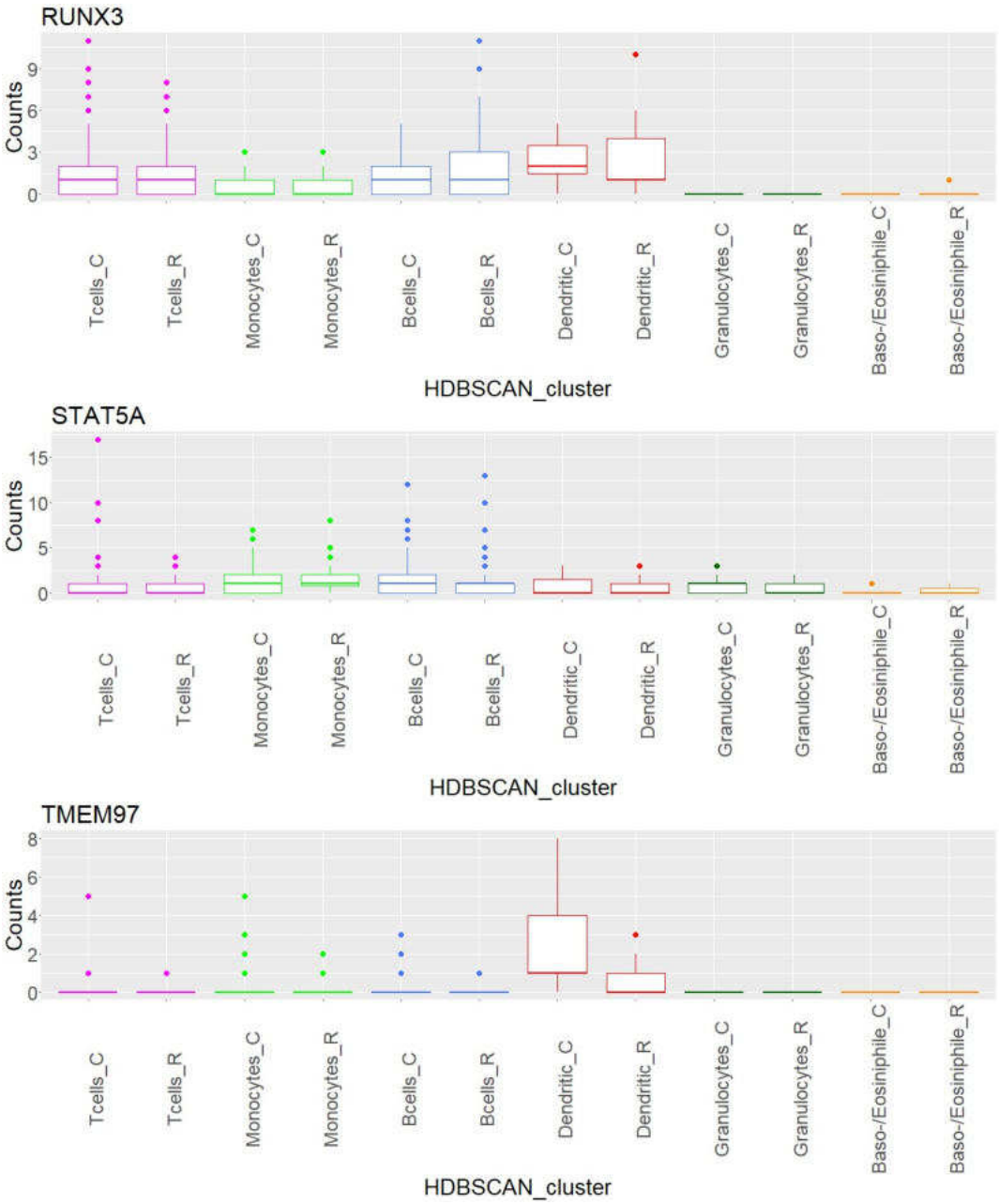

CXCL9
Counts
150
100
50
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster
DDB2
Counts
10.0
7.5
5.0
2.5
0.0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster
FLT3
Counts
8
6
4
2
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster

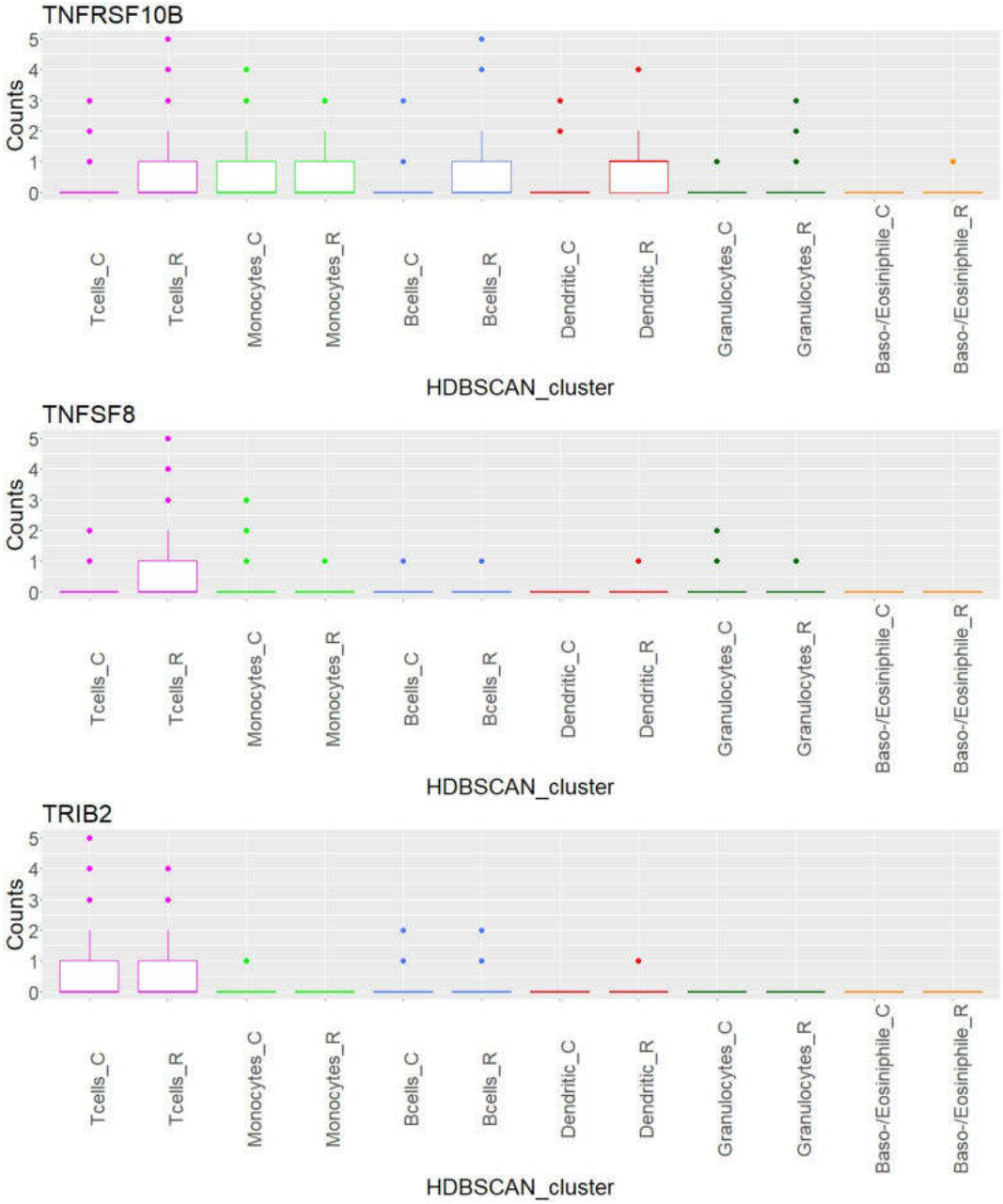

FYB
Counts
15
10
5
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster
GAPDH
Counts
80
60
40
20
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster
GZMH
Counts
15
10
5
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster

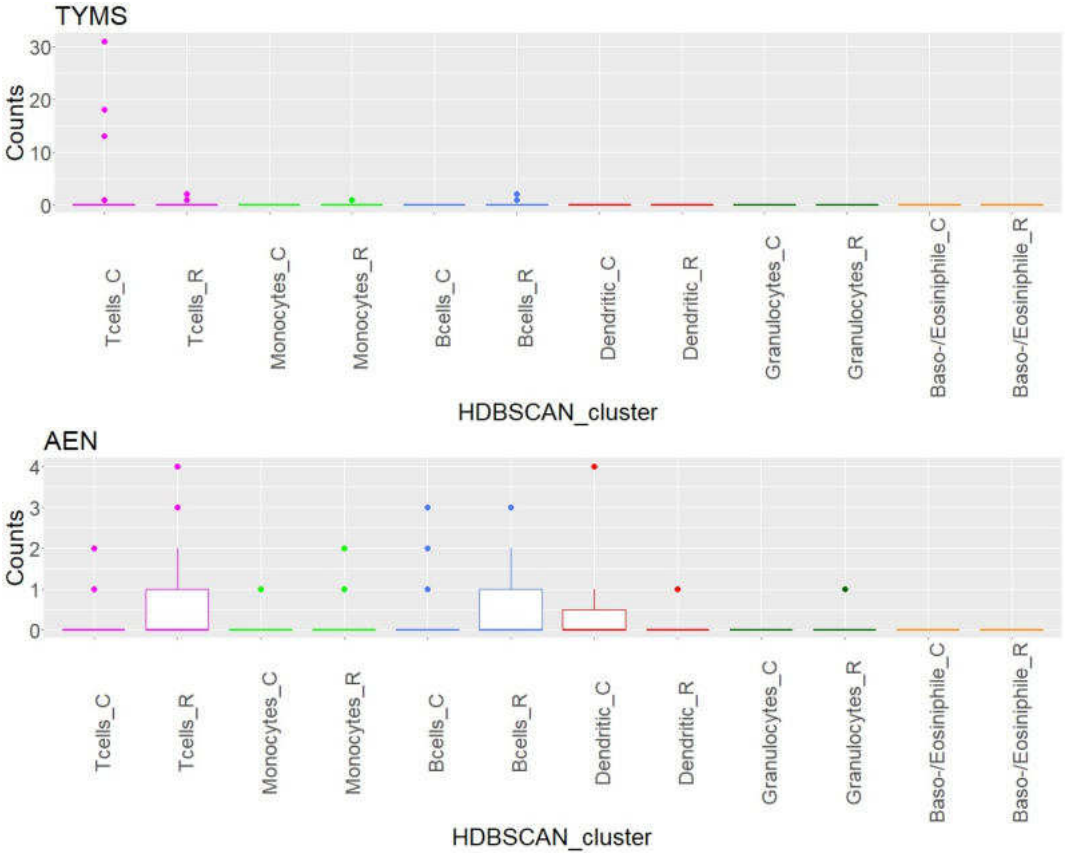

IL15
Counts
HDBSCAN_cluster
LAT
Counts
HDBSCAN_cluster
LCK
Counts
HDBSCAN_cluster
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R


MCM2
Counts
20
15
10
5
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster
MYC
Counts
60
40
20
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster
PHPT1
Counts
15
10
5
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster


RPL23AP42
Counts
150
100
50
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster
RPS19P1
Counts
150
100
50
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster
RPS27L
Counts
30
20
10
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster


RUNX3
Counts
9
6
3
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster
STAT5A
Counts
15
10
5
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster
TMEM97
Counts
8
6
4
2
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster


TNFRSF10B
Counts
0
1
2
3
4
5
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster
TNFSF8
Counts
HDBSCAN_cluster
TRIB2
Counts
HDBSCAN_cluster


TYMS
Counts
30
20
10
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster
AEN
Counts
4
3
2
1
0
Tcells_C
Tcells_R
Monocytes_C
Monocytes_R
Bcells_C
Bcells_R
Dendritic_C
Dendritic_R
Granulocytes_C
Granulocytes_R
Baso-/Eosiniphile_C
Baso-/Eosiniphile_R
HDBSCAN_cluster

Ex vivo data functional analysis

Table 40. Functional analysis based on ex vivo data results with marked 91 statistically significant BPs.

Description	Gene Ratio	pvalue	p.adjust	geneID
intrinsic apoptotic signaling pathway by p53 class mediator	5/18	1,13007E-08	1,27472E-05	RPS27L/AEN/BAX/CD74/ CDKN1A
intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator	4/18	1,26284E-07	7,12244E-05	RPS27L/AEN/CD74/ CDKN1A
intrinsic apoptotic signaling pathway	6/18	2,2756E-07	8,55626E-05	LCK/RPS27L/AEN/BAX/ CD74/CDKN1A
cellular response to UV	4/18	8,02119E-07	0,00022383	DDB2/BAX/CDKN1A/MYC
regulation of cysteine-type endopeptidase activity involved in the apoptotic process	5/18	1,29299E-06	0,00022383	LCK/THBS1/RPS27L/BAX/ MYC
regulation of fibroblast proliferation	4/18	1,33066E-06	0,00022383	BAX/CD74/CDKN1A/MYC
fibroblast proliferation	4/18	1,38902E-06	0,00022383	BAX/CD74/CDKN1A/MYC
regulation of cysteine-type endopeptidase activity	5/18	2,65348E-06	0,000304757	LCK/THBS1/RPS27L/BAX/ MYC
intrinsic apoptotic signaling pathway in response to DNA damage	4/18	2,70174E-06	0,000304757	RPS27L/AEN/CD74/ CDKN1A
DNA damage response, signal transduction by p53 class mediator	4/18	2,70174E-06	0,000304757	RPS27L/BAX/CD74/CDKN1A
cellular response to light stimulus	4/18	3,81074E-06	0,000390774	DDB2/BAX/CDKN1A/MYC
signal transduction by p53 class mediator	5/18	5,3783E-06	0,000505561	RPS27L/AEN/BAX/CD74/CDKN1A
signal transduction in response to DNA damage	4/18	6,0638E-06	0,000526152	RPS27L/BAX/CD74/CDKN1A
positive regulation of cysteine-type endopeptidase activity involved in the apoptotic process	4/18	6,80256E-06	0,000548092	LCK/RPS27L/BAX/MYC
response to UV	4/18	7,81721E-06	0,000587854	DDB2/BAX/CDKN1A/MYC
response to ionizing radiation	4/18	9,1777E-06	0,000647028	AEN/BAX/CDKN1A/MYC
positive regulation of cysteine-type endopeptidase activity	4/18	1,09775E-05	0,000728391	LCK/RPS27L/BAX/MYC
positive regulation of fibroblast proliferation	3/18	1,35869E-05	0,000851445	CD74/CDKN1A/MYC
positive regulation of proteolysis	5/18	1,7745E-05	0,00102561	LCK/RPS27L/BAX/TRIB2/MYC
response to gamma radiation	3/18	1,89054E-05	0,00102561	BAX/CDKN1A/MYC
regulation of epithelial cell proliferation	5/18	1,98614E-05	0,00102561	THBS1/STAT5A/BAX/STAT1/MYC
cellular response to radiation	4/18	2,0003E-05	0,00102561	DDB2/BAX/CDKN1A/MYC
positive regulation of endopeptidase activity	4/18	2,26888E-05	0,001112737	LCK/RPS27L/BAX/MYC
mitotic G1 DNA damage checkpoint	3/18	2,79097E-05	0,001259284	RPS27L/BAX/CDKN1A
mitotic G1/S transition checkpoint	3/18	2,79097E-05	0,001259284	RPS27L/BAX/CDKN1A
G1 DNA damage checkpoint	3/18	2,92013E-05	0,001266888	RPS27L/BAX/CDKN1A
positive regulation of peptidase activity	4/18	3,29575E-05	0,001316904	LCK/RPS27L/BAX/MYC
regulation of endopeptidase activity	5/18	3,32093E-05	0,001316904	LCK/THBS1/RPS27L/ BAX/MYC
response to interleukin-9	2/18	3,5024E-05	0,001316904	STAT5A/STAT1
cell proliferation involved in metanephros development	2/18	3,5024E-05	0,001316904	STAT1/MYC
epithelial cell proliferation	5/18	3,86638E-05	0,001406863	THBS1/STAT5A/BAX/ STAT1/MYC
B cell proliferation	3/18	4,09741E-05	0,001444337	BAX/CD74/CDKN1A

Description	Gene Ratio	pvalue	p.adjust	geneID
regulation of dendritic cell antigen processing and presentation	2/18	4,27841E-05	0,001462438	THBS1/CD74
regulation of peptidase activity	5/18	4,47916E-05	0,001486028	LCK/THBS1/RPS27L/BAX/MYC
response to radiation	5/18	4,81257E-05	0,001551022	DDB2/AEN/BAX/CDKN1A/MYC
dendritic cell antigen processing and presentation	2/18	5,13132E-05	0,001607815	THBS1/CD74
cell cycle arrest	4/18	5,93145E-05	0,00180829	THBS1/BAX/CDKN1A/MYC
activation of cysteine-type endopeptidase activity involved in apoptotic process	3/18	6,17124E-05	0,001831884	LCK/RPS27L/BAX
metanephric mesenchyme development	2/18	8,15028E-05	0,002357311	STAT1/MYC
G1/S transition of mitotic cell cycle	4/18	0,000104144	0,002807809	RPS27L/BAX/CDKN1A/MYC
DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator	2/18	0,000105452	0,002807809	RPS27L/CDKN1A
mitotic DNA damage checkpoint	3/18	0,000105617	0,002807809	RPS27L/BAX/CDKN1A
regulation of myeloid cell differentiation	4/18	0,000107035	0,002807809	THBS1/STAT1/CD74/MYC
mitotic DNA integrity checkpoint	3/18	0,000118268	0,002972137	RPS27L/BAX/CDKN1A
DNA damage response, signal transduction resulting in transcription	2/18	0,000118569	0,002972137	RPS27L/CDKN1A
regulation of B cell apoptotic process	2/18	0,000132447	0,003149614	BAX/CD74
kidney mesenchyme development	2/18	0,000132447	0,003149614	STAT1/MYC
cell cycle G1/S phase transition	4/18	0,000134026	0,003149614	RPS27L/BAX/CDKN1A/MYC
lymphocyte proliferation	4/18	0,000157772	0,003593657	BAX/CD74/CDKN1A/TNFSF8
mononuclear cell proliferation	4/18	0,000161679	0,003593657	BAX/CD74/CDKN1A/TNFSF8
cell proliferation involved in kidney development	2/18	0,000162479	0,003593657	STAT1/MYC
cellular response to abiotic stimulus	4/18	0,000178012	0,003731399	DDB2/BAX/CDKN1A/MYC
cellular response to environmental stimulus	4/18	0,000178012	0,003731399	DDB2/BAX/CDKN1A/MYC
regulation of monocyte differentiation	2/18	0,000178631	0,003731399	CD74/MYC
cellular response to tumor necrosis factor	4/18	0,000186612	0,003827252	CHI3L1/THBS1/STAT1/TNFSF8
response to light stimulus	4/18	0,000193257	0,003892758	DDB2/BAX/CDKN1A/MYC
negative regulation of G1/S transition of mitotic cell cycle	3/18	0,000205662	0,004069944	RPS27L/BAX/CDKN1A
leukocyte proliferation	4/18	0,000219095	0,004261026	BAX/CD74/CDKN1A/TNFSF8
B cell apoptotic process	2/18	0,000231613	0,004348481	BAX/CD74
regulation of metanephros development	2/18	0,000231613	0,004348481	STAT1/MYC
negative regulation of cell cycle G1/S phase transition	3/18	0,000235157	0,004348481	RPS27L/BAX/CDKN1A
response to tumor necrosis factor	4/18	0,000247346	0,004490141	CHI3L1/THBS1/STAT1/TNFSF8
positive regulation of mesenchymal cell proliferation	2/18	0,000250779	0,004490141	STAT1/MYC
B cell homeostasis	2/18	0,000291361	0,005135242	BAX/CD74
positive regulation of leukocyte activation	4/18	0,000321222	0,00557444	LCK/THBS1/CD74/CDKN1A
DNA damage checkpoint	3/18	0,000373307	0,006380155	RPS27L/BAX/CDKN1A
positive regulation of cell activation	4/18	0,000398352	0,006706591	LCK/THBS1/CD74/CDKN1A

Description	Gene Ratio	pvalue	p.adjust	geneID
regulation of antigen processing and presentation	2/18	0,000431021	0,00695112	THBS1/CD74
regulation of mesenchymal cell proliferation	2/18	0,000431021	0,00695112	STAT1/MYC
DNA integrity checkpoint	3/18	0,000431364	0,00695112	RPS27L/BAX/CDKN1A
mitotic cell cycle checkpoint	3/18	0,000446732	0,007097382	RPS27L/BAX/CDKN1A
monocyte differentiation	2/18	0,000510867	0,007800349	CD74/MYC
mononuclear cell differentiation	2/18	0,000510867	0,007800349	CD74/MYC
regulation of intrinsic apoptotic signaling pathway	3/18	0,000511725	0,007800349	LCK/BAX/CD74
regulation of endothelial cell proliferation	3/18	0,000546379	0,008217536	THBS1/STAT5A/STAT1
negative regulation of fibroblast proliferation	2/18	0,000567785	0,008370137	BAX/MYC
regulation of apoptotic signaling pathway	4/18	0,000581233	0,008370137	LCK/THBS1/BAX/CD74
positive regulation of protein kinase B signaling	3/18	0,000582506	0,008370137	LCK/CHI3L1/THBS1
negative regulation of protein phosphorylation	4/18	0,000586206	0,008370137	BAX/TRIB2/CDKN1A/ MYC
regulation of G1/S transition of mitotic cell cycle	3/18	0,000610583	0,008609221	RPS27L/BAX/CDKN1A
positive regulation of B cell proliferation	2/18	0,000627642	0,008740495	CD74/CDKN1A
regulation of T cell receptor signaling pathway	2/18	0,000658669	0,009060717	LCK/PHPT1
positive regulation of apoptotic signaling pathway	3/18	0,000679437	0,009233788	LCK/THBS1/BAX
endothelial cell proliferation	3/18	0,00068966	0,009261152	THBS1/STAT5A/STAT1
regulation of smooth muscle cell proliferation	3/18	0,000710402	0,009427453	THBS1/STAT1/CDKN1A
smooth muscle cell proliferation	3/18	0,000731539	0,009595071	THBS1/STAT1/CDKN1A
mesenchymal cell proliferation	2/18	0,000756132	0,009692237	STAT1/MYC
regulation of granulocyte chemotaxis	2/18	0,000756132	0,009692237	THBS1/CD74
response to mechanical stimulus	3/18	0,000786131	0,009883667	CHI3L1/THBS1/STAT1
myeloid cell differentiation	4/18	0,00079248	0,009883667	THBS1/STAT1/CD74/ MYC
regulation of cell cycle G1/S phase transition	3/18	0,000797353	0,009883667	RPS27L/BAX/CDKN1A