Politecnico di Torino

POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering

Academic year 2023/2034

Master's Degree Thesis

AI-driven morphological clustering for lymphoma stratification

Supervisors:

Ing. Massimo SALVI

Prof. Filippo MOLINARI

Candidate:

Isotta MELONI

Table of contents

List of tables *2*

List of figures *3*

List of abbreviations *5*

Abstract *6*

Introduction *8*

1.1 Lymphoma **8**
1.1.2 Pathological anatomy 10
1.1.3 DLBCL morphology 11
1.1.4 DLBCL gene expression profile 12
1.1.5 Multi-Omnic Analysis 14

1.2 Artificial intelligence in medicine **18**
1.2.1 Machine Learning 19
1.2.2 Deep Learning 20

1.3 Clustering for patient stratification **23**
1.3.1 Unsupervised Clustering 23
1.3.1.1 K-means 24
1.3.1.2 Hierarchical Clustering 25

Methods *26*

2.1 Description of the dataset **26**

2.2 Pipelines **27**

2.3 Data Preparation **28**
2.3.1 Generation of binary masks 28
2.3.2 Processing Masks 29
2.3.3. Tile Processing 31

2.4 Feature extraction **37**
2.4.1 Features extracted from binary masks 40
2.4.1.1 Graphs 42
2.4.2 Features extracted from histological tiles 44

2.5 Patient clustering and survival analysis **46**

Results *49*

3.1 Analysis of the population of subjects **49**

3.2 Comparison of clustering algorithms **50**
3.2.1 Splitting into two clusters 50
3.2.2 Division into three clusters 55
3.2.3 Division into four clusters 60

3.3 Critical issues and improvements **65**

Conclusion and future developments *67*

Bibliography *69*

List of tables

Table 1 – Major events associated with the patient: death status, any first event, time any first event..... 48
Table 2 - Comparison of the clusters parameters (k=2)..... 55
Table 3- Comparison of the clusters parameters (k=3)..... 59
Table 4 - Comparison of cluster parameters (k=4) 64

List of figures

Figure 1 - Digitization of histological images....10
Figure 2 - Common morphological variants of DLBCL in H&E: (A) centroblastic variant; (B) immunoblastic variant; (C) anaplastic variant [1] 12
Figure 3 – DLBCL classes by genetic distinction....13
Figure 4 - Multi-omnic approaches....15
Figure 5 - Main methods of integration of multi-omics data according to Picard....18
Figure 6 - AI layering....19
Figure 7 - ML Learning Techniques 20
Figure 8 - Difference Between Machine Learning and Deep Learning....21
Figure 9 - Types of learning for clinical image processing 21
Figure 10 - Relationship between the amount of data and the performance achieved in machine learning and deep learning 22
Figure 11- Example of histological ROI (1024x1204) with H&E staining extracted on the patient (21_L_1503_A1_EE_tile_18)....26
Figures12- Process flow diagram....27
Figure 13 - (A) ROI extracted; (B) result of the softmax algorithm used; (C) overlapping the main boundaries with the original ROI; (D) binary mask obtained. (Patient reference: 22_L_565_A1A_EE_tile_25) 28
Figure 14 - Flowchart of the mask processing algorithm used 30
Figure 15 - First phase of mask processing: reduction of false segmentation 30
Figure 16 - (A) Original tile; (B) overlapping of the processed mask on the original tile (cores identified in yellow, cores discarded in blue)....31
Figure 17 - Tile preprocessing flowchart....32
Figure 18 - Result obtained from the processing of the tiles 33
Figure 19 - Optimized tile pre processing....35
Figure 20 - Result of the application of optimized tile processing 36
Figure 21- In the green box the optimized and saved tiles; The other images are examples of discarded tiles. The cards refer to different groups of patients, specified in the figure....37
Figure 22 - Manual selection of DLBCL cell diameter extremes: left end selection, right end selection. 38
Figure 23 - (A) mask of the nuclei; (B) identified cell populations: in yellow the lymphocytes, in purple the DLBCL cells 39
Figure 24 - Feature extraction flowchart....40
Figure 25 - Flowchart of features extracted from binary masks of nuclei....41
Figure 26 - Superposition of centroids (A) and graph (B) on the binary mask of nuclei 43
Figure 27 - Example of a graph obtained from the population of small cells....43
Figure 28 - Possible offset directions [23]....44
Figure 29 - Flowchart of the extraction of texture features extracted from histopathological tiles....45
Figure 30 - (A) histological tile; (B) the edges identified in tile A are highlighted in yellow....46

Figure 31 - Occurrence in months since the occurrence of the first adverse event 49
Figure 32 - The result of the grouping obtained with the k-means algorithm (k=2), with the number of deaths, displayed by the first two main components. 50
Figure 33 - Silhouette graph for the k-means algorithm (k=2) 51
Figure 34 - Adverse event survival curve for the k-means algorithm (k=2) 52
Figure 35 - Result of the grouping obtained with the linkage algorithm (k=2), with number of deaths, displayed by means of the first two main components. 52
Figure 36 - Dendrogram representative of hierarchical clustering obtained from normalized features (k=2) ... 53
Figure 37- Silhouette graph for the linking algorithm (k=2) 53
Figure 38 - Result of the grouping obtained with the linkage algorithm (k=2), with number of deaths, displayed by the first two main components 54
Figure 39 - The result of the grouping obtained with the K-means algorithm (k=3), with the number of deaths, displayed by means of the first two main components. 55
Figure 40 - Silhouette graph for the K-means algorithm (k=3) 55
Figure 41 - Adverse event survival curve for the K-means algorithm (k=3) 56
Figure 42 - Result of the grouping obtained with the linkage algorithm (k=3), with number of deaths, displayed by means of the first two main components. 57
Figure 43 - Dendrogram representative of hierarchical clustering obtained from normalized features (k=3) ... 57
Figure 44 - Silhouette graph for linkage algorithm (k=3) 58
Figure 45 - Result of the grouping obtained with the linkage algorithm (k=3), with number of deaths, displayed by means of the first two main components 59
Figure 46 - The result of the grouping obtained with the K-means algorithm (k=4), with the number of deaths, displayed by the first two main components. 60
Figure 47 - Silhouette graph for the K-means algorithm (k=4) 60
Figure 48 - Adverse event survival curve for the K-means algorithm (k=3) 61
Figure 49 - Result of the grouping obtained with the linkage algorithm (k=4), with number of deaths, displayed by means of the first two main components. 62
Figure 50 - Dendrogram representative of hierarchical clustering obtained from normalized features (k=4) ... 62
Figure 51 - Silhouette graph for the K-means algorithm (k=4) 63
Figure 52 - Result of the grouping obtained with the linkage algorithm (k=4), with number of deaths, displayed by means of the first two main components 63

List of abbreviations

Abbreviation	Meaning	Pages
DLBCL	Diffuse large B-cell lymphoma	6
HL	Hodgkin lymphoma	8
NHL	Non-Hodgkin lymphoma	8
NOS	Not otherwise specified	8
GEP	Gene expression profiling	8
R-CHOP	rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone	9
AI	Artificial intelligence	9
WSI	Whole slide imaging	10
H&E/EE	Hematoxylin and eosin	12
CGB	Diffuse large B-cell lymphoma similar to germ center B-cells	13
ABC	Diffuse large B-cell lymphoma with activated B-cells	13
COO	Cells of origin	13
CNV	Copy number variation	16
SNV	Simple nucleotide variation	16
ML	Machine learning	19
DL	Deep Learning	20
RGB	Red, green and blue color pattern	26
HSV	Hue saturation value	34
PCA	Principal Component Analysis	47
AFE	Any first event	47
NaN	Not a number	48

Abstract

Diffuse large B-cell lymphoma (DLBCL) is the most common lymphoid neoplasm in Western countries and represents an aggressive form of lymphoma, particularly heterogeneous in terms of symptoms, genetic profile and therapeutic response. This pathology presents significant challenges for researchers and clinicians due to its considerable variability.

The heterogeneous nature of DLBCL, in fact, makes its management complex and difficult to identify subgroups of patients who may benefit from specific treatments. The variability in symptoms, genetic characteristics and response to therapies complicates the analysis and requires a multidisciplinary and integrated approach.

One of the main obstacles to research is the difficulty in finding complete and accurate data on patients with DLBCL. Furthermore, morphological data are often insufficient to provide a complete picture of the disease. It is crucial, therefore, to collect a wide range of information, including not only morphological, but also genetic, molecular, clinical and immunological data.

Using AI algorithms to manage vast data sets is critical to understanding the complexity of the disease and tumor environment to improve survival predictions in cancer studies.

The stratification of cancer patients represents a crucial step in the personalization of therapies and the optimization of clinical approaches, allowing a separation of patients into homogeneous groups.

The objective of this study is the clustering of DLBCL patients, as a function of morphological characteristics, using artificial intelligence to obtain a stratification of subjects that reflects their survival. Histopathological images and clinical data of patients with DLBCL were used to obtain homogeneous groups related to prognosis, using unsupervised partition algorithms. The use of hierarchical clustering has allowed the obtaining of two subgroups of populations, different in morphological characteristics and probability of survival to adverse events, such as disease aggravation or recurrence. The two groups also reflect the percentage of associated deaths and therefore the prognosis of the subjects themselves.

The integration of these machine learning methods into clinical practice can lead to the optimization of treatment strategies and better disease management, increasing the chances of therapeutic success and improving the quality of life of patients.

It is therefore essential to intensify the commitment to data collection to improve the understanding of DLBCL and refine patient stratification techniques.

Introduction

1.1 Lymphoma

Lymphoma is a neoplasm that originates in the lymphatic system, a key component of the immune system. This disease is divided into two main categories: Hodgkin lymphoma (HL) and non-Hodgkin lymphoma (NHL), each characterized by different subtypes, pathogenesis, and clinical behaviors.

Diffuse large B-cell lymphoma is the most common lymphoid neoplasm in Western countries and accounts for 30-40% of non-Hodgkin's lymphoma (NHL) cases. 85% of cases are known as "not otherwise specified" (NOS) and represent an aggressive, particularly heterogeneous tumor, in terms of symptoms, genetic profile and therapeutic response. [1]

DLBCL is characterized by rapid growth of B lymphocytes, one of two types of white blood cells. The disease derives, in fact, from the neoplastic transformation of lymphoid cells and is genetically associated with a vast number of mutations, chromosomal translocations and epigenetic alterations.

The diagnosis is in a wide age range, with a median localized around 64 years, and its course depends mainly on the extra-lymph node site in which it occurs. [2]

The detection of the pathology requires a multidisciplinary and highly specialized approach to the patient, which usually combines the use of histological investigations, clinical analysis and advanced molecular techniques. The complex and heterogeneous network of information related to this subtype of lymphoma strongly influences the prognosis and the choice of therapeutic treatment.

Characterized by a 5-year overall survival rate of 60-70%, the best approach for DLBCL characterization is personalized medicine. The use of gene expression profiling (GEP) studies allows, in fact, the optimization of the therapy, designed on the specific patient, thanks to the identification of target genes. [3] [4] [5]

The treatment of lymphomas varies greatly depending on the subtype, stage, and molecular characteristics of the tumor. In non-Hodgkin's lymphomas, the therapeutic

approach can range from active surveillance in indolent cases to more aggressive chemotherapy regimens and targeted therapies. Standard treatment involves the use of R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone). The response rate sees 25% of patients immune to therapy or subject to relapses, requiring salvage therapy with high-dose immuno-chemotherapy and/or allogeneic transplantation.

Over the past 20 years, several attempts have been made to try to classify DLBCL-NOS and better predict its therapeutic response.

This was made possible thanks to the progressive development of high-throughput technologies that have greatly increased knowledge of the molecular characteristics and oncogenic mechanisms responsible for the development and progression of DLBCL. [6] However, despite the introduction of state-of-the-art tools and recent advances in the study of lymphoma, the use of such systems within clinical practice is still impractical.

This is because, although a limited number of biomarkers have been identified for the characterization of the molecular subtype, these are still poorly reproducible and are associated with poor prognostic value. [6]

Lymphoma represents a complex and dynamic field of cancer research. Within the diverse and heterogeneous system that characterizes lymphoma, artificial intelligence (AI) can play an important role in identifying characteristics or associations hidden from the human eye, useful for identifying any prognostic biomarkers, outlining the elements that differentiate the various neoplastic subtypes.

The use of artificial intelligence (AI)-based algorithms could be easily integrated into pathology laboratories equipped with slide scanners, to increase current diagnostic/clinical performance.

The aim of the following study is to identify the features that characterize subjects with DLBCL and that influence their survival status.

The correct stratification of patients makes it possible to identify homogeneous groups, different from each other in terms of characteristics and course of the disease. In this way, it is possible to analyze the characteristics that unite the subjects of the same cluster and identify optimal therapeutic pathways, targeted and personalized on the patient, which would lead to an improvement in the prognosis of the subjects themselves. This research stems from a growing clinical need and the need to increase the reliability of the therapies used, to reduce the number of subjects immune to treatment and the number of relapses.

The analysis conducted aims to create an algorithm capable of differentiating subjects affected by DLBCL based on the intrinsic relationships between morphological data and the

outcome of the disease, to support the pathologist in the decision-making process and the medical team in defining personalized paths and treatments for patients.

1.1.2 Pathological anatomy

Pathological anatomy is a discipline aimed at studying morphological, immunophenotypic and molecular alterations that involve tissues in different pathological processes, such as infections, inflammation and neoplasms.

Histological examination is a diagnostic medical procedure that involves the microscopic analysis of tissue samples, taken for the study of pathologies.

Histology is a branch of biology that deals with the study of cells and their organization within a tissue, both from a morphological and functional point of view.

To allow their study, the tissue must be fixed, cut and colored with colors functional to the recognition of the various components that make up a tissue.

Digital pathology is a field of pathology that allows the digitization of histological images taken from the patient during biopsy, using high-resolution scanners (Fig.1). The digital images produced by the scanners can be viewed with the latest generation software and analyzed through automatic processes, facilitating the management, sharing and manipulation of complex data, in a highly reproducible way.

Modern slide scanners enable fast, high-resolution digitization of large portions of tissues (Whole-Slide Imaging – WSI), thus making available an enormous amount of data that reflects the morphological and functional aspects (in the case of immunohistochemical staining) of tissues. [6] This method allows the excised tissue section to be quickly visualized and analyzed, generating a digital file that can be easily used within complex computational algorithms.

Figure 1 - Digitization of histological images

Automating the analysis of digitized histology slides is a fast, high-throughput, and cost-effective alternative to manual analysis, characterized by slowness and possible errors

related to inter/intra-operator variability. The use of artificial intelligence in medicine offers vast possibilities for improving diagnostic and clinical processes in cancer treatment.
In particular, artificial intelligence can certainly play a fundamental role in speeding up the extraction and management of the large amount of data associated with patients, providing concrete help to clinicians.
The information extracted, together with the knowledge and experience of the pathologist, outlines the diagnosis and the best therapeutic path for the patient analyzed.

In oncology, pathological anatomy is essential for determining the type of tumor, the stage of the disease, and for determining possible therapeutic targets.
In clinical practice, histopathological imaging analysis is based on the interpretation of the morphological and chromatic characteristics of the observed slides.

Different operators then lead to different interpretations of the same slide. Pathological analysis is, in fact, strongly linked to the experience of the operator and influenced by it.
There is, therefore, an evident variability in the quality of the service that can be provided to the citizen.
The adoption of digital algorithms in pathology could, therefore, help reduce much of this variability, standardizing the medical procedure and bringing it to a higher quality.
The digital transition to the disease, enhanced by artificial intelligence, could provide benefits to all patients. [6]

1.1.3 DLBCL morphology

DLBCL is a disease morphologically characterized by tumor cells that are larger than the benign cells present within the same portion of tissue, usually larger even than tissue macrophages. DLBCL cells are round or ovoid in shape, presenting a diffuse neoplastic growth, which invades the lymphatic tissue.

The most common variants described are the centroblastic variant, the immunoblastic variant and the anaplastic variant. The characteristics of the variants are:

A. centroblastic variant: In this variant, the cancer cells resemble centroblastic cells, immature B cells involved in the production of antibodies. It represents the most widespread variant, covering about 80% of cases;
B. immunoblastic variant: in this variant, most of the cells (90%) are made up of immunoblasts, activated mature B cells that release antibodies into the body;

C. anaplastic variant: this variant is characterized by the presence of degenerated tumor cells, which lose their resemblance to normal lymphoid cells; It represents the least frequent variant, covering about 3% of all cases.

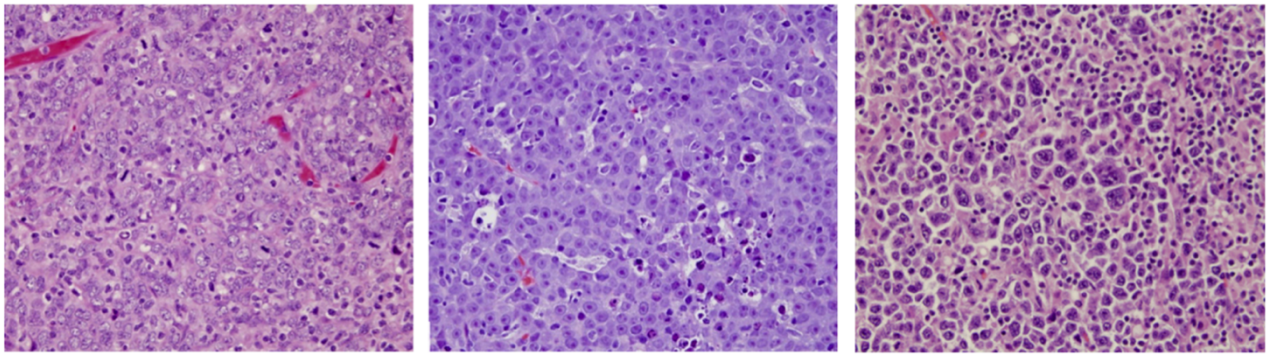


Figure 2 - Common morphological variants of DLBCL in H&E: (A) centroblastic variant; (B) immunoblastic variant; (C) anaplastic variant [1]

Despite the different histological configurations, the morphology of DLBCL is not currently correlated with the prognosis of subjects. To this end, the use of automated image processing algorithms could identify potential biological and immunohistochemical markers at the prognostic level. [7]

1.1.4 DLBCL gene expression profile

Cancerous diseases are often characterized by alterations in gene expression. Certain genes, known as oncogenes, can be overexpressed, promoting, for example, the uncontrolled growth of cancer cells. Others may be responsible for the evasion of cancer cells from the normal mechanisms of apoptosis. The gene expression profile is a quantitative analysis that allows to compare genes expressed differently in a pathological tissue with those expressed by a healthy tissue, to understand the mechanisms involved in the pathology analyzed and to allow their study. The use of gene expression profiling allowed DLBCL to be classified into three categories:

- CGB: subgroups of diffuse large B-cell lymphoma similar to germ center B-cells;
- ABC: Activated B-cell Diffuse Large B-cell Lymphoma Subgroups
- unclassified;

The identified subgroups have different stages of differentiation and activation of B cells, also known as "cells of origin" (COO). The various categories are associated with different clinical

outcomes, with GCB patients associated with significantly higher overall survival and free survival than ABC patients. [6]

Many studies have focused on defining predictive models for clustering genetic subtypes of DLBCL, exploiting the onset of shared genetic abnormalities.

A 2018 study [8] identified 5 distinct classes of DLBCL. The study uses an automated method that starts from a series of classes (seeds) and iteratively moves cases in and out of classes to obtain a method for genetic distinction. The identified phenotypic subtypes, shown in yellow in the figure below, differ in gene expression signatures and responses to immunotherapy.

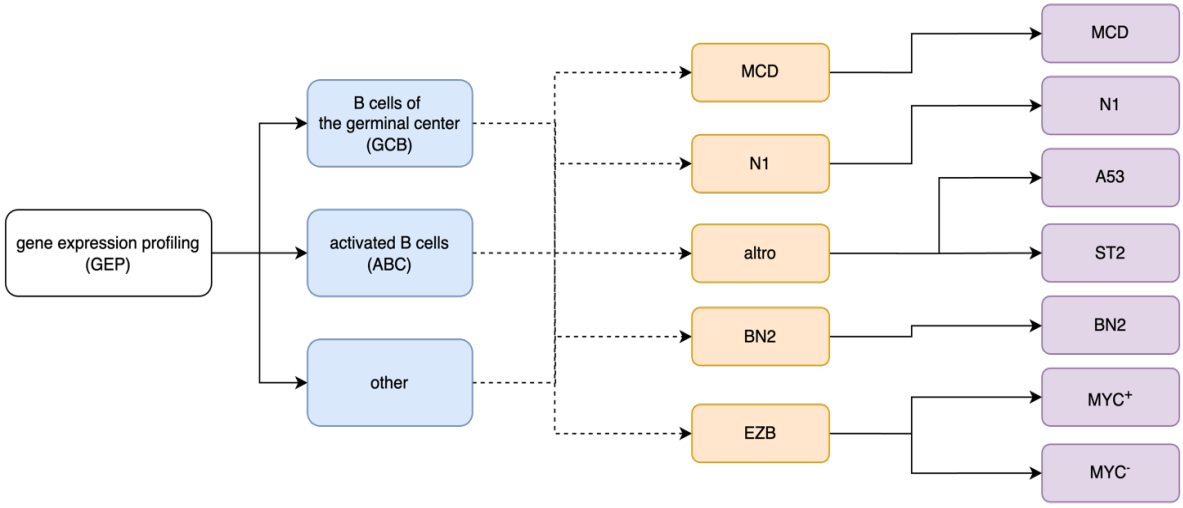


Figure 3 – DLBCL classes by genetic distinction

A subsequent study [9], based on a Bayesian predictive model, expands the number of previously identified genetic signatures and determines the probability of lymphoma belonging to one of the seven identified classes, each of which is associated with a biomarker (in purple).

Although algorithms based on immunohistochemical evaluation have been identified for the identification of biomarkers, these are numerically very limited. The reproducibility and prognostic and predictive role of DLBCL-associated biomarkers, although generally included in the pathology report, is still controversial [6].

There is therefore a strong need to develop robust classification methods, aimed at defining the mechanisms and pathways that drive the development and progression of DLBCL, emphasizing the need to develop specific therapies. [10]

There are various disciplines aimed at studying information derived from alteration in the number of copies, chromosomal arrangements, epigenetic alterations and gene expression of subjects. The information derived from genetic profiles is numerous and requires elaborate analysis techniques to identify targets of prognostic value associated with diseases.
In the study of complex pathologies, as in the case of DLBCL, the availability of such information is however difficult, due to the high heterogeneity between subjects which turns into a difficult comparative analysis.

1.1.5 Multi-Omnic Analysis

Multi-omnic analysis consists of the integrated use of genetic data and clinical data within algorithms for predicting survival, for analyzing the biological processes of oncological diseases and personalizing therapies.

The analysis of omnic data, collected at scale, requires the use of advanced methods for the identification of significant patterns in the biological data and processes analyzed.

The goal is to understand which molecular components interact and contribute to the functioning of a complex biological process.

Understanding the biological processes involved in cancer diseases is a fundamental step for biomedical research, to develop innovative, targeted and personalized therapies on the individual subject, overcoming the "one-size-fits-all" approach.

Multi-omnic analysis generally increases the accuracy of prediction in survival analyses by plotting the relationships between biological mechanisms and consequences on the body of the analyzed subject, as a clinician would.

The term "multi-omnic" refers to the study of the genome, transcriptome, epigenome, proteome, exposome, and microbiome.

Below is a graphical diagram representing an overview of the main analyses of omnimic data (Fig. 4).

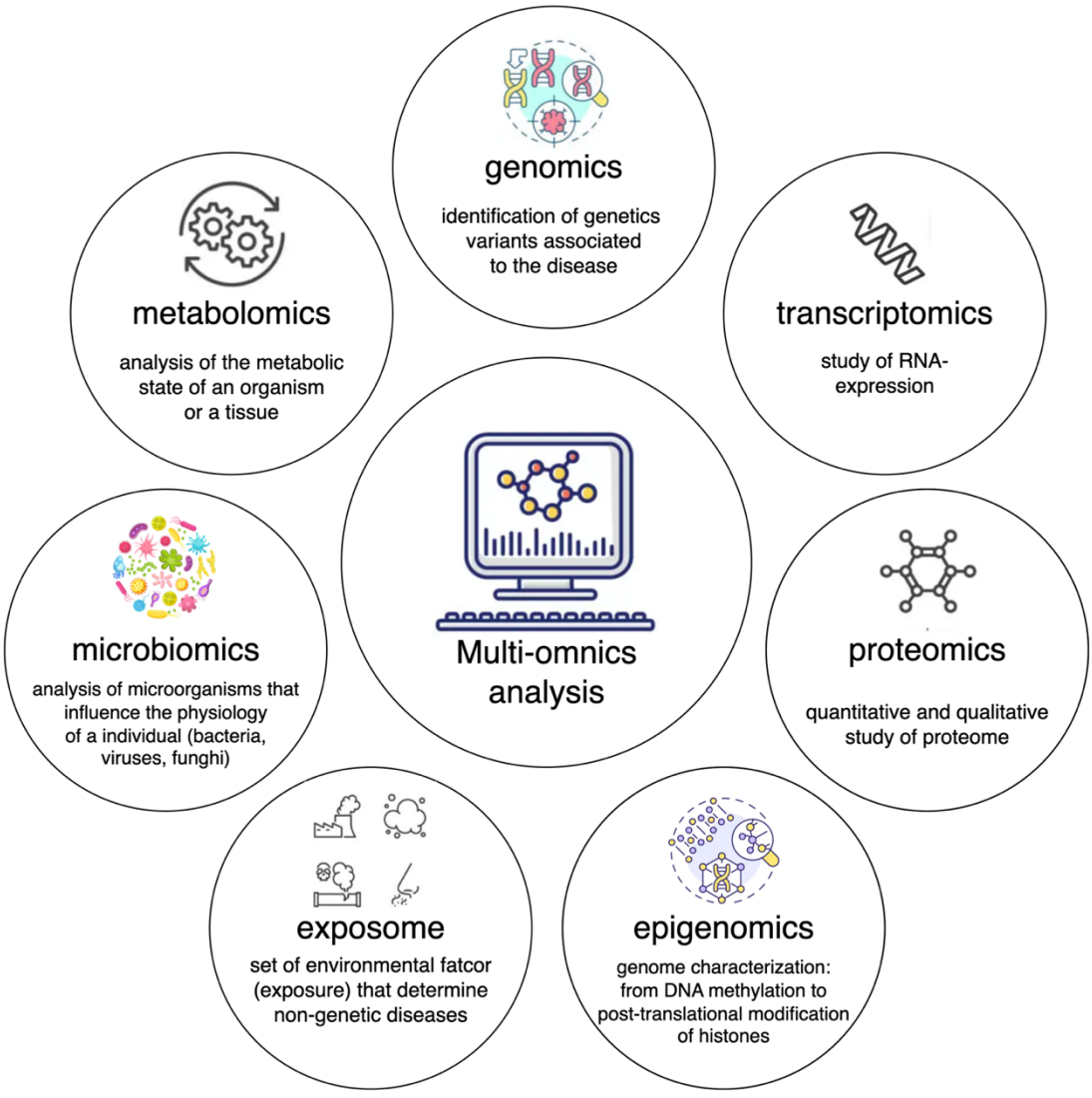


Figure 4 - Multi-omnic approaches

The term genomics encompasses all methods of analyzing the sequence and structure of DNA to identify genetic variants that may be associated with specific medical conditions. The use of DNA-microarrays, sensors that allow the simultaneous analysis of various genetic portions, has made it possible to analyze DNA on a large scale, revolutionizing branches such

as medicine and biology, increasing the effectiveness of therapies through personalized medicine [9].

Copy number variation (CNV), together with simple nucleotide variation (SNV), are methods of analysis that fall under genomics.

NVC describes the phenomenon of the repetition of genomic sequences, characteristic of all. Such repeats may represent biomarkers in tumor disease processes.

SNV is the process of replacing a single nucleotide within the DNA strand which, depending on its location within a coding or non-coding region, can lead to the formation of a pathogenic variant or the premature truncation of a protein. SNV leads to slight variations in the genome, associated with individual diversity and predisposition to certain diseases or response to external agents [9].

These phenomena affect gene expression and, consequently, the protein expression of the individual. [11]

The transcriptome profile analyzes the complete set of all RNA molecules present within a cell or tissue at any given time.

The aim of the analysis is to map both coding and non-coding components, determining the heterogeneity of gene expression within cells, organs and tissues. [12]

The analysis allows to reconstruct the networks of biological interaction, producing a molecular fingerprint of the processes.

The proteome profile (the set of proteins synthesized by the genome) represents the large-scale study of the proteins expressed by an organism at a given time. The study of the proteome focuses on the analysis of post-translational changes and protein-protein interactions [9].

The proteome analysis allows, in the study of DLBCL, to stratify patients into different subcategories; This stratification, however, does not correspond to the prognostic outcome of the stratified subjects.

The analysis conducted in the following study aims to identify a stratification of patients based on morphological features derived from the subjects' histopathological slides, so that clustering corresponds to the prognostic output of the subjects' survival.

Unlike the genome, both the transcriptome and the proteome are highly dynamic entities, which do not remain constant over time.

Epigenomics is a branch of molecular biology that studies chemical changes at the epigenetic level that can affect gene expression and how these changes can be passed on to subsequent generations. The most common modifications concern DNA methylation, a

regulatory process of cell differentiation and repression, and post-translational histone alterations, a process related to the mechanisms of transcription or gene inhibition. [13]

The exposome analyzes the exposure to environmental factors to which the individual is exposed and the influence that these factors can have on the body, investigating the relationships between external agents and the development of diseases. This branch allows the exploration of factors external to the body.

Microbiomics analyzes the microorganisms that influence the physiology of an individual. The microbiota is made up of bacteria, archaea, viruses, phages and fungi. This branch studies how changes in microbiome activity and composition affect individuals' disease states. [14]

Finally, metabolomics studies the quantification and expression of metabolites in a biological sample, focusing on the variation of the metabolic profile in various biological settings [13].

Multi-omnic sciences allow the understanding of biological mechanisms, studying the interactions between internal and external factors, allowing the identification of potential biomarkers or the identification of pharmacological targets, with consequent improvement of diagnostic-therapeutic pathways.

The main challenges in analyzing large data sets in multi-omnics integration [13] are:

- dimensionality curse: This is about the amount of data collected about a single subject in relation to contextually limiting the subjects on whom the data is taken. This aspect leads to several problems, including overfitting, the phenomenon of hyper-adaptation to training data with loss of the ability to generalize, the increase in computational cost and the increase in noise. These elements contribute to increasing the difficulty of interpreting the results and reducing the robustness of the analysis method used;
- data heterogeneity: This refers to the inherent diversity of data and the methods used to analyse it.

Multi-omnic analysis requires the use of integration methods that reduce system complexity and computational time, which allow for the optimal selection of the models that most correlate with the desired output.

The main methods of multi-omnic data integration, theorized by Picard, are graphically shown in the figure below. [17]

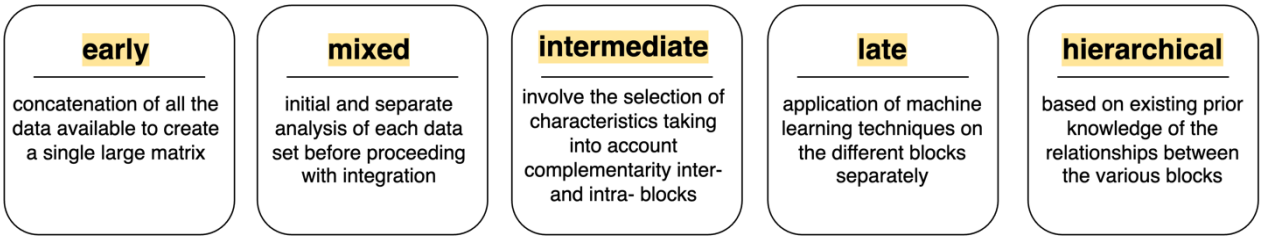


Figure 5 - Main methods of integration of multi-omics data according to Picard

Data integration methods provide an overview of the most used ways to process data and how they are used together within analytics pipelines. [15] [16]
Early and mixed integration strategies are the easiest to implement as they involve concatenating data into a single matrix, respectively, or initially analyzing several datasets separately before integrating them. These methods, although performing, ignore the intrinsic distribution of data and the possible complementarity within the various omics blocks [17].
Intermediate integration involves the selection of features that share a common latent space, to simulate molecular interactions between biological mechanisms. The main advantage of these techniques is their ability to identify the inter-omic structure of the joint by emphasizing the complementary information in each omics [17].
Late integration involves the application of machine learning methods on the different omics blocks, to combine their respective predictions (acronym-omnics multiple analysis).
Finally, hierarchical integration strategies use external information, from scientific databases, sequentially on the different data, exploiting previous knowledge on the interactions between the various blocks.

Advanced genetic testing can be particularly expensive when performed for the analysis of rare or complex diseases. Furthermore, the availability of data could be hindered by the use of different protocols or machinery present in the various research centers, by the lack of homogeneity of the data and by the different storage format..

1.2 Artificial intelligence in medicine

Artificial intelligence (AI) is a field of data science that aims to create algorithms that can simulate human intelligence. AI encompasses several approaches that differ depending on the computational potential used (Fig. 6).

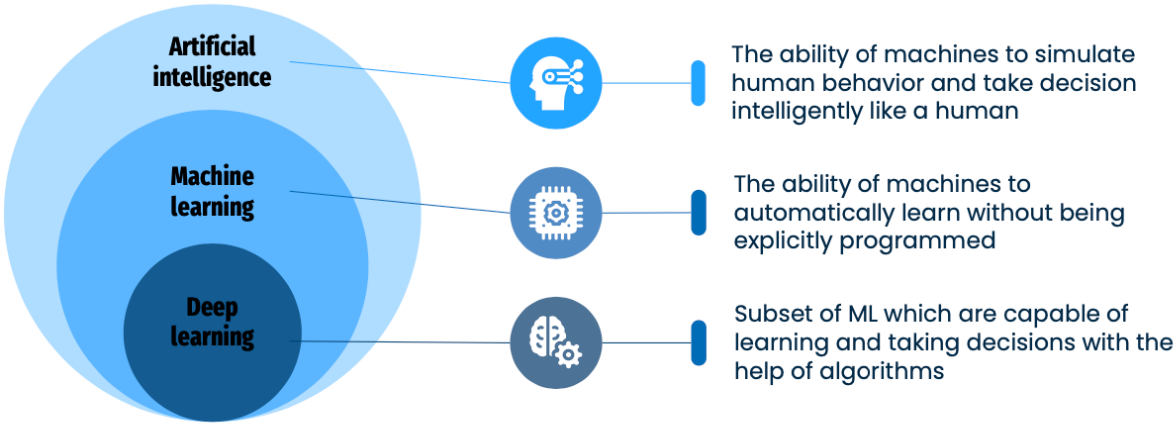


Figure 6 - AI layering

Artificial intelligence offers new horizons to current clinical practices, providing methods to improve the diagnosis, treatment and management of diseases. Machine learning algorithms can analyze large amounts of genetic, imaging and clinical data, providing diagnostic support tools, which could revolutionize the therapeutic and clinical approach. Using automated algorithms in medicine can increase efficiency, reduce errors, and make care targeted and personalized.

1.2.1 Machine Learning

Machine learning (ML) encompasses learning methods that can automatically derive insights from data without the need to program code to do so. This technology is based on learning algorithms that can analyze and interpret large volumes of data, to identify patterns and trends that allow predictions or decisions to be made, leveraging predetermined patterns and/or equations.

As the model learns, it improves its performance based on the number of examples made available in the system. The operation of the algorithms requires that the model takes as input the features, manually extracted from the data, learns the internal connections and produces a result. Machine learning is divided into three different types, depending on the learning modalities (Fig.7), each useful for solving different tasks [18]:

1. supervised learning: Use models from the training dataset to map out the characteristics of the target, using the information learned to make predictions about future data. During training, the model compares the predictions obtained with the labels provided and updates the model to minimize the error made on the prediction. It is used for regression or classification tasks;

2. unsupervised learning: Use unlabeled data to investigate patterns hidden in the data itself, without external supervision. It is used for clustering tasks, to group instances into separate clusters, based on specific combinations of the characteristics themselves;
3. reinforcement learning: use both the information learned from the data and the information learned during the mistakes made during the training phase, to solve a specific task. It is therefore based on a cycle of actions and feedback, used to optimize the long-term strategy, taking advantage of the external dynamic environment.

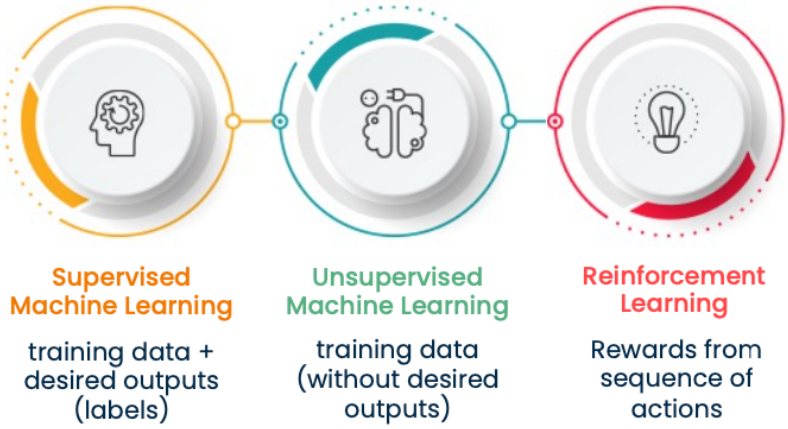


Figure 7 - ML Learning Techniques

1.2.2 Deep Learning

Deep learning (DL) is an evolution of machine learning that uses interconnected deep networks to model complex relationships between data. It is particularly useful when managing large datasets, both structured and unstructured.

The algorithms used in DL have multiple layers of processing, linked together to form a hierarchy of features. Each layer of the network takes the information learned from the previous layer and processes it further, allowing the automatic extraction of high-level features.

The main difference between DL and ML is that in deep learning, the features are not chosen directly by the user but by the algorithm itself, within the learning process (Fig. 8).

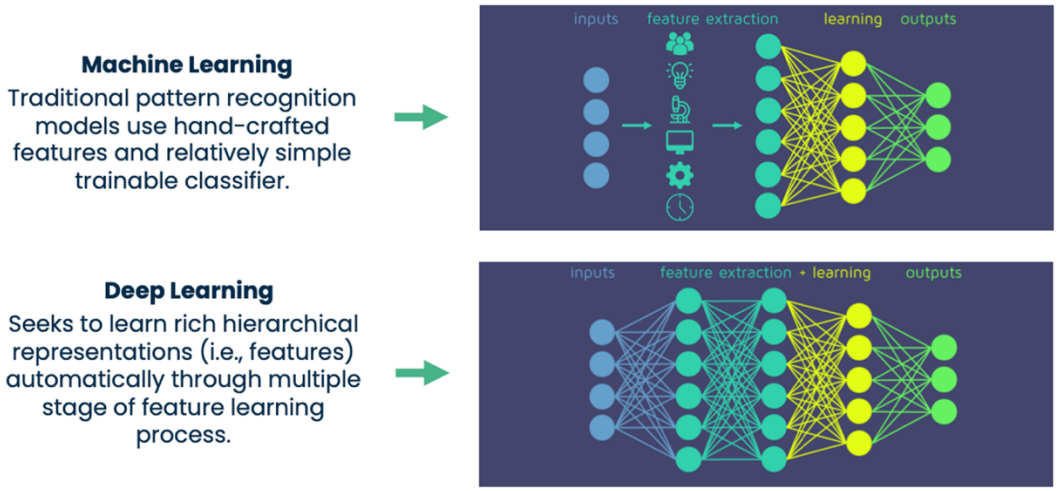


Figure 8 - Difference Between Machine Learning and Deep Learning

The types of learning differ depending on the purpose. In clinical image processing, the main ones are divided into classification, detection, segmentation, and generation (Fig. 9).

- Classification allows you to assign a label or class to a dataset, using predefined categories. This method can be used for the classification and grading of a given pathology;
- Detection is the process of identifying and locating a target of interest within a frame or image. It is used to detect any tissue injuries or degeneration;
- Segmentation is the process of classifying pixels in an image, to classify them into a certain class. The process makes it possible to segment and separate pathological regions or entire organs, to analyze their contents;
- Generation allows the creation of new data, starting from reference images. The goal is to expand the information on which to apply subsequent learning algorithms.

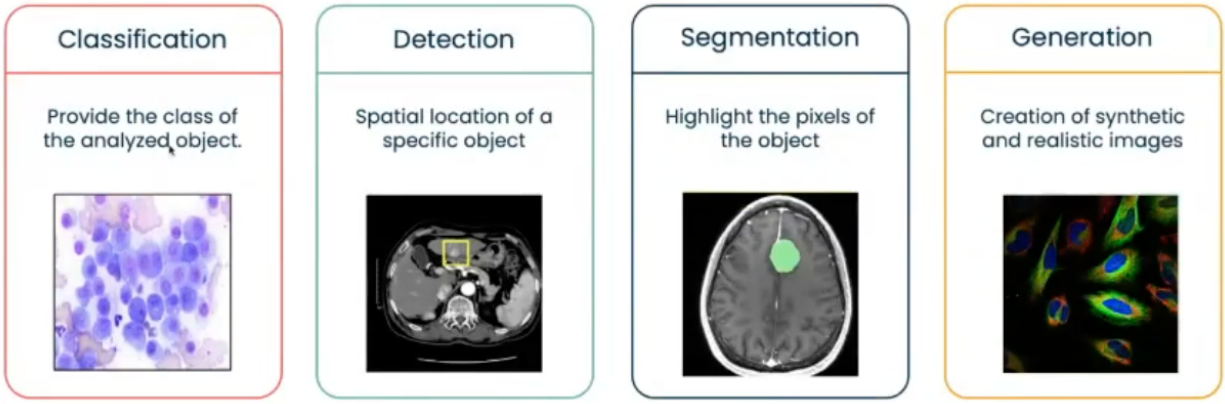


Figure 9 - Types of learning for clinical image processing

The performance you get using an ML or DL algorithm depends on the amount of data you use. ML algorithms tend to improve quickly at first as the amount of data provided increases. However, this trend tends to saturate, indicating that a significant increase in data does not also lead to improved performance. The advantage of using these methods lies in a lower computational cost, due to a lower complexity of the methods themselves and of the problem under consideration and a shorter associated resolution time. DL algorithms, on the other hand, tend to have lower initial performance that improves as the amount of data provided increases. The improvements are, in fact, continuous and more consistent than the ML (Fig. 10).

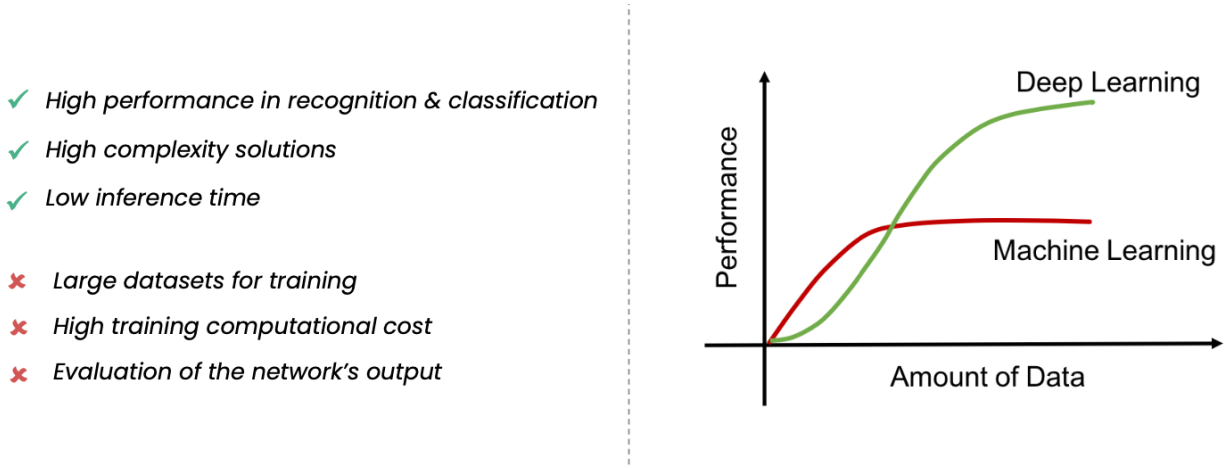


Figure 10 - Relationship between the amount of data and the performance achieved in machine learning and deep learning

The use of clinical, histological and genetic data within predictive models can be exploited in oncology to estimate patient survival and the degree of aggressiveness of the disease. The use of learning algorithms in cancer research is an important tool for biomedicine, offering numerous benefits ranging from early detection to the discovery of new biological targets. These benefits come from the ability to analyze significant amounts of complex data and identify hidden patterns that are difficult to detect with traditional methods.

Cancer is a complex disease that requires the integrated analysis of data of different kinds and the collaboration of multidisciplinary teams. By simulating the practice followed by a doctor, AI algorithms can be used to automatically integrate and analyze different types of characteristics.

Automated systems produce output by significantly reducing both the times and costs associated with traditional diagnostic methods and the workload of healthcare professionals, simultaneously increasing the scalability of diagnostic and therapeutic processes and optimizing the reproducibility of results. [19]

1.3 Clustering for patient stratification

The use of clustering algorithms can be useful for the subdivision of cancer patients into homogeneous subgroups, which are difficult to identify with traditional approaches. Heterogeneity among cancer patients is one of the biggest challenges in cancer care. In fact, even if tumors are of the same type, they can behave differently and lead to different outcomes in patients who apparently have similar characteristics.
Clustering in machine learning allows you to automatically manage the heterogeneity of data and process it effectively, creating subgroups with significant differences in treatment response, prognosis, or disease progression.

1.3.1 Unsupervised Clustering

Unsupervised clustering is a machine learning method that uses unlabeled data to generate groups of data (or clusters), without any prior knowledge of the category to which they belong.
In clustering, the goal is to identify, within the analyzed data, hidden patterns or structures, generating homogeneous groups. Within each cluster, the data have common characteristics that are all the more distant from the other clusters the better the method used, in relation to the data provided to the system itself.

A clustering algorithm is characterized by:

- a grouping rule;
- a measure of similarity.

To assess the similarity between the elements, a "prototype" is used; The prototype is an element, not necessarily real, that characterizes the elements of a cluster as a group.

In the biomedical field, unsupervised clustering methods are particularly useful in the analysis of large heterogeneous datasets, to identify hidden patterns that allow patient stratification and therefore the optimization of treatments and the personalization of care pathways. The ability to discover new, non-obvious relationships between data is critical for clustering patients, identifying new biomarkers, and predicting disease progression over time. Unsupervised learning techniques tend to outperform supervised methods in the analysis of relatively small datasets, typical of biomedical databases. [20]

In the study of a complex disease such as large B-cell lymphoma, it is essential to identify strategies for a significant stratification of subjects.

The study carried out proposes the analysis and comparison of two different types of clustering, in order to allow a subdivision of patients that reflects the analysis of their survival and their prognosis The unsupervised clustering methods used are K-means and hierarchical clustering, which differ in their ability to handle different types of data, similarity parameter and sensitivity to initial parameters.

1.3.1.1 K-means

K-means is a partitional, unsupervised clustering algorithm used to generate a defined K number of disjoint groups.

Each cluster in the K-means is associated with a representative point (prototype), called a centroid, usually calculated as the average of the elements. The number of centroids calculated is equal to the number K of groups to be formed.

Additional parameters that characterize clusters are intra-cluster variability and inter-cluster variability. Intra-cluster variability is a measure of similarity between elements belonging to a cluster, such as distance between points in the same cluster and its centroid, while inter-cluster variability measures the distance between the groups themselves.

During the initial phase of data processing, centroids are randomly assigned. Each element is then iteratively assigned to the nearest cluster, leading to the computation of a new centroid. The iterative process ends when one or more conditions are met:

- the centroids stabilize: the cluster assignments for the individual points no longer change and the algorithm converges towards the solution;
- the algorithm has completed executing the specified number of iterations.

K-means is generally used in the presence of continuous data sets, as it is based on data distance parameters. On discrete or categorical data sets, in fact, the algorithm may not lead to meaningful solutions.

The most common measure of similarity used is the Euclidean distance (d) between the data. The purpose of the algorithm is, in fact, to minimize the squared distance between the data and the center of gravity of the cluster to which they belong.

$$d(x, c) = (x - c)(x - c)'$$

This measure makes the algorithm highly sensitive to the scale of variables and the presence of outliers, assigning greater weight to data with a higher numerical value. To obtain results that are independent of variable scale, you must perform normalization methods on the data, so that the result is not biased.

The result obtained by clustering also depends heavily on the K parameter, which refers to the number of clusters into which the dataset is intended to be partitioned. The a priori definition of a parameter is difficult when preliminary information about the data is not known. The a priori definition of the K parameter can lead to the achievement of only sub-optimal results.

1.3.1.2 Hierarchical Clustering

Hierarchical clustering produces a hierarchy of nested clusters. The simplest representation of hierarchical clustering is a tree-like pattern, known as a dendrogram. In the dendrogram, each cluster union is represented by a branch, the height of which indicates the distance or dissimilarity between the groups. The length of the segment that joins the elements within the clusters, in fact, is proportional to the degree of dissimilarity.
In hierarchical clustering, you don't need to define the number of clusters in advance. This makes it easier to use, especially for datasets for which no prior structural information is known.
In the bottom-up approach, the elements are considered as a single cluster, consisting of maximum similarity, to be iteratively joined to neighboring elements, each represented by a prototype. The merging of the individual elements continues until a single cluster is obtained.
In the less used top-down approach, you start with a single cluster containing all the data, and progressively divide it into clusters contained by individual elements.
Again, the quality of the clusters obtained can depend heavily on the similarity measure used.

In the context of agglomerative hierarchical clustering methods, one of the methods that can be used for group construction is '*Ward'*. This is used for the creation of compact and homogeneous clusters, with minimal internal variance. Unlike other methods, the *'eard'* tends to avoid the formation of elongated or irregularly shaped clusters.

The use of hierarchical clustering techniques presents a great flexibility, thanks to the a posteriori choice of the number of groups to be formed and to a graphic and intuitive visualization of the similarity between the elements provided by the dendrogram.

Methods

2.1 Description of the dataset

The data analyzed relate to a total of 117 subjects suffering from diffuse large B-cell lymphoma. Histopathological images of the solid tumor in RGB, data on the status of death of the subject and temporal data relating to the presence of a first adverse event, starting from diagnosis, were provided of the available patients.

For each patient, 50 40x histological ROIs, measuring 1024x1024, were extracted from digitized histology slides. ROIs were randomly derived from the original WSI, without the application of a specific algorithm or method to identify areas of interest. This is because, since it is a solid tumor, it was hypothesized that each area of the tile could constitute an area of clinical interest.

The analyzed images are stained with hematoxylin eosin (H&E). This stain is used in histopathology to highlight pathological changes present within an organ or tissue. Inside the slide, the cytoplasm and its basic components are stained pink, while the nucleus and various acidic components are stained blue-violet (Fig. 11).

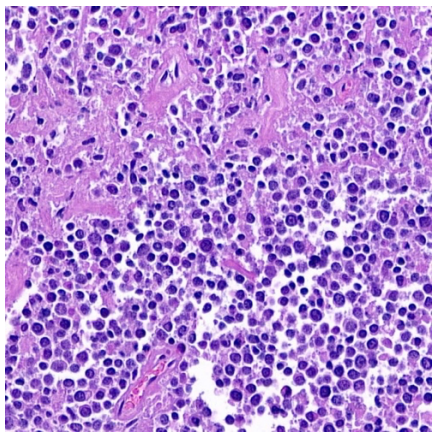


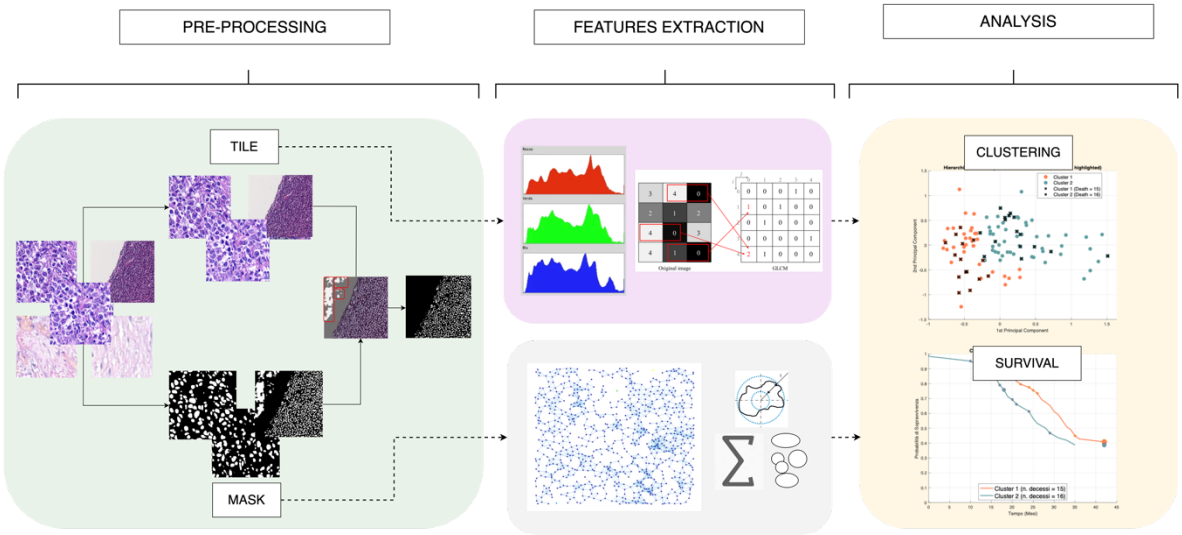
Figure 11- Example of histological ROI (1024x1204) with H&E staining extracted on the patient (21_L_1503_A1_EE_tile_18)

Histological data (ROI) were pre-processed to reduce areas of non-clinical interest (example: areas with high slide content). Binary masks relating to the distribution of tissue nuclei were then extracted from the processed tiles.
From the data thus obtained, the features that characterize each subject were extracted, used as input to the tested algorithms.

2.2 Pipelines

The aim of the study is the separation of the analyzed subjects into clusters that reflect the probability of survival and/or the state of death of the subjects themselves, using automatic partitioning algorithms.

In the first phase of the project, histological data were processed to obtain tissue ROIs and the binary masks associated with them. The data obtained were then processed to create a reduced and homogeneous dataset, which reduced the areas of non-diagnostic interest present in the tiles and excluded the histological regions consisting of artifacts. On the data obtained, the characterizing characteristics were then extracted, both on the histological tiles and on the binary masks, to identify the morphological parameters that represented the individual subjects. The matrix of features thus obtained was used within two clustering algorithms, K-means and hierarchical clustering, to compare their performance and identify the most appropriate method for the separation of subjects based on their morphological characteristics. Finally, the survival curves of the event were estimated on the clusters obtained, to verify the goodness of the clusters based on the prognosis of the subject.



Figures12- Process flow diagram

The entire analysis process, shown in figure 12, has been developed and processed in *Matlab*.

2.3 Data Preparation

The histological tiles obtained were processed to obtain a reduced dataset of increased quality. The reduced dataset was used in the later stages of the analysis.

The processing process for increasing the quality of imaging data concerns:

- use of a three-class network for the generation of binary masks, related to the nuclei scheme;
- processing of the masks obtained;
- tile processing.

The processes will be analyzed in detail in the following chapters.

2.3.1 Generation of binary masks

The masks of the nuclei were automatically generated using an automatic algorithm for the discrimination of three classes.

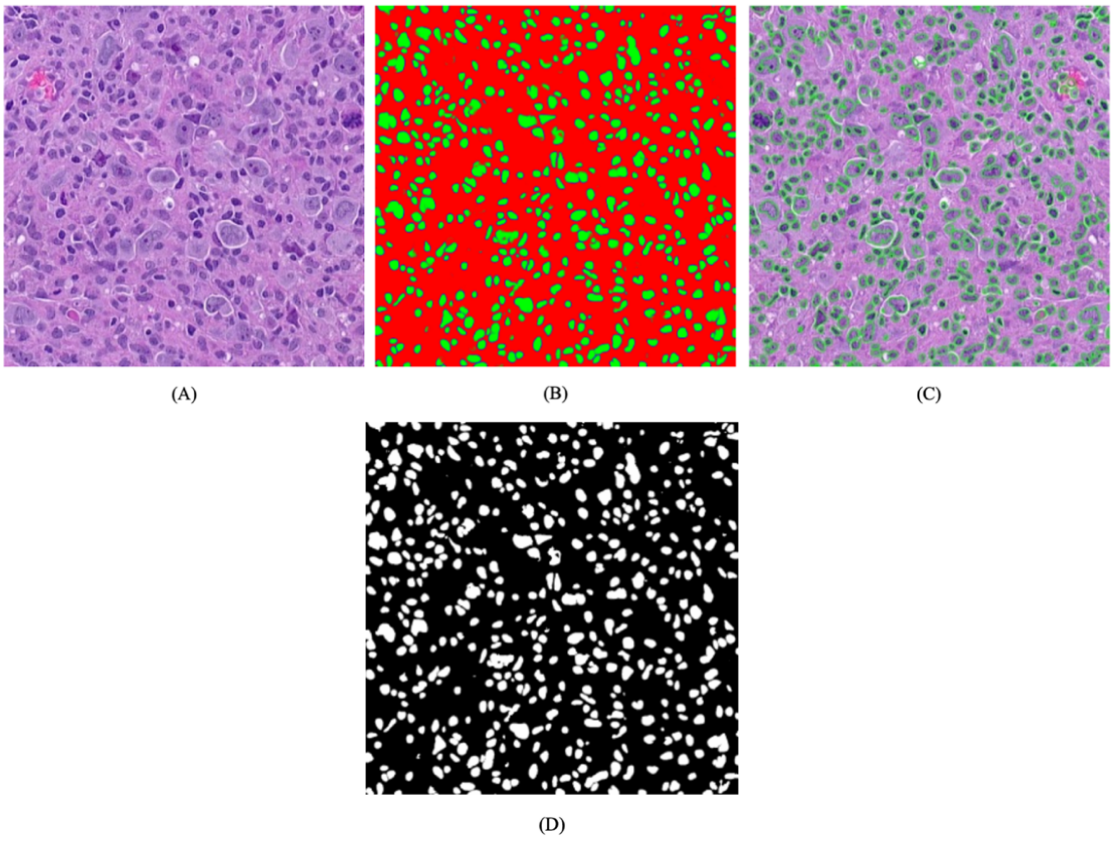


Figure 13 - (A) ROI extracted; (B) result of the softmax algorithm used; (C) overlapping the main boundaries with the original ROI; (D) binary mask obtained. (Patient reference: 22_L_565_A1A_EE_tile_25)

This method allows to obtain probability masks (softmax), starting from the histological images colored in H&E. Softmax allows you to interpret the parts in the image as the probability of belonging to a class. There are three possible classes: cell nucleus, nucleus contours, and background (cytoplasm).

Once the probabilities are known, binary masks of segmented nuclei can be obtained, without merging neighboring objects. The outputs obtained from the network are shown in Figure 13.

Obtaining accurate cell segmentation is a necessary requirement for computational analysis of histological data. In the clinical field, in fact, the information obtained from cell morphology is fundamental within the diagnostic-decision-making process. [7]

2.3.2 Processing Masks

Once the binary masks were obtained, they were processed to allow a correct management of the data obtained.

In the initial phase, masks were superimposed on the tiles to evaluate, nucleus by nucleus, the segmentation obtained by the three-class network.

For each segmented nucleus, the average intensity of the color associated with it was evaluated, to evaluate whether the segmented area leaving the network corresponded to a cell nucleus.

Areas segmented with the "core" class from the network correspond to regions of tissue colored purple. This, using the histological grayscale image as a reference, corresponds to values of low intensity (dark color).

In this way, the nuclei corresponding to the areas of high intensity (white color) were identified, to remove them.

Figure 14 shows the flow chart of the process used.

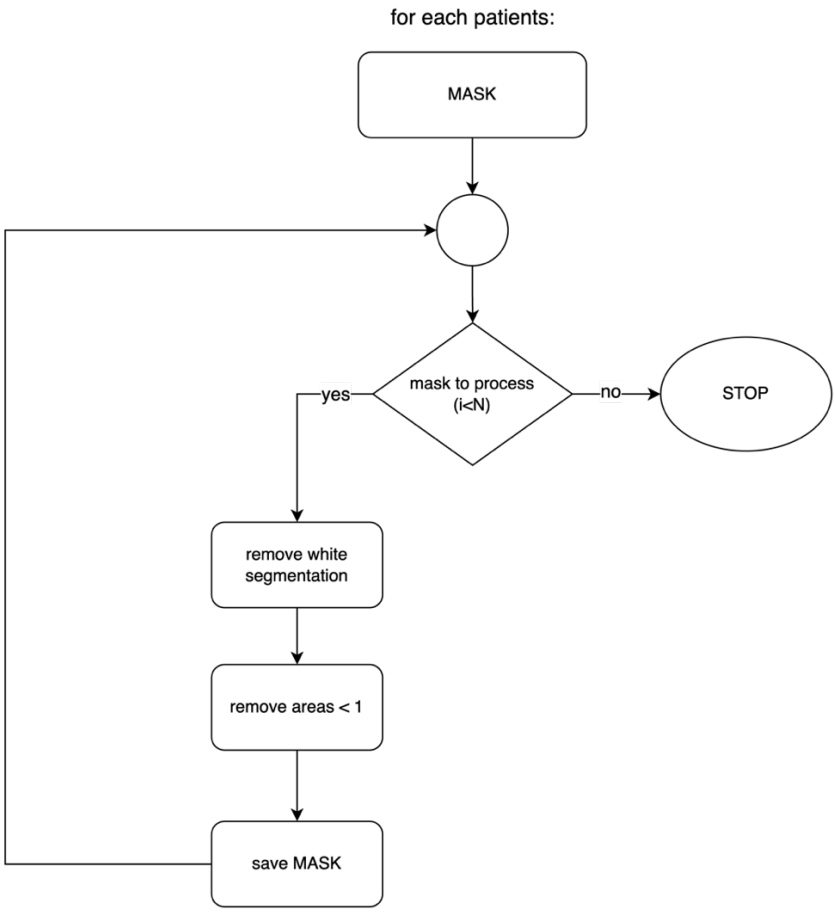


Figure 14 - Flowchart of the mask processing algorithm used

This reduces areas of false segmentation that do not correspond to the nucleus of a cell but to the histological glass. This process is shown graphically in figure 15.

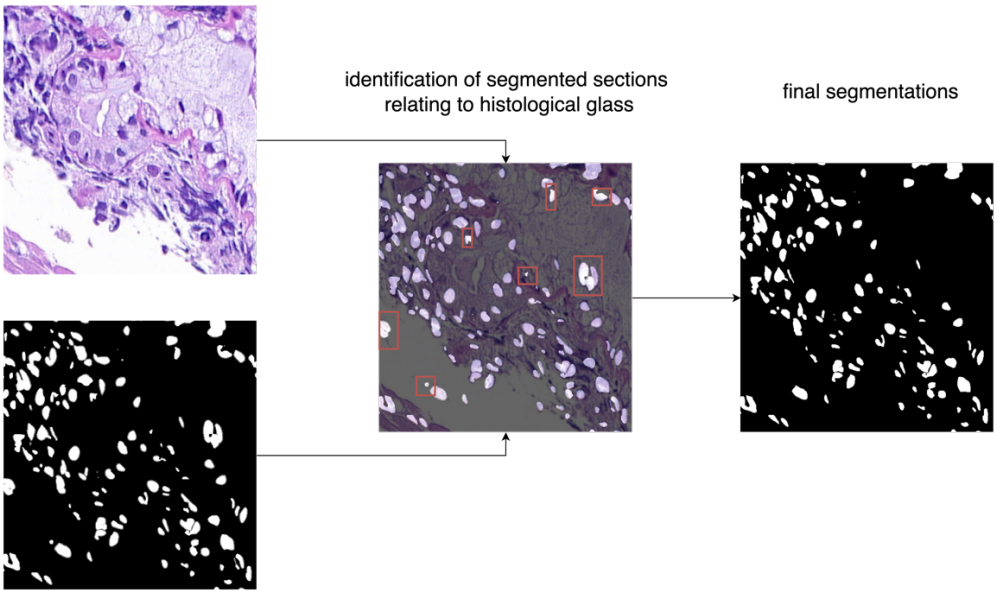
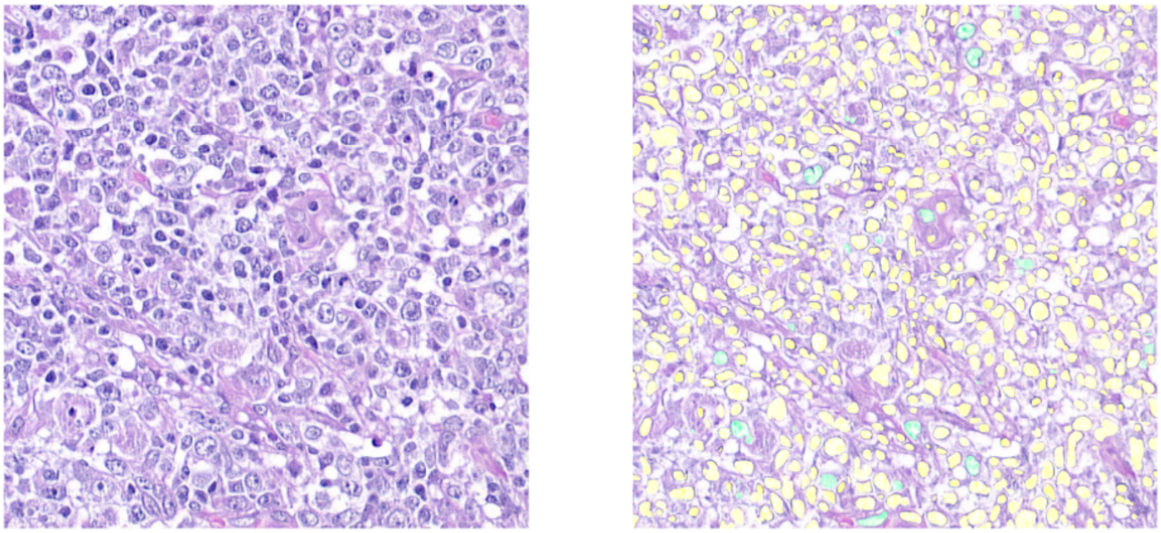


Figure 15 - First phase of mask processing: reduction of false segmentation

An automatic algorithm was then applied to the output masks to remove the cores that had an irrelevant area (zero or unitary). This process is necessary for the correct generation of the graphic files and for the extraction of the morphological characteristics used within the algorithms used for the analysis of the survival of the subjects. The importance of this step was verified during the feature extraction phase.

The result of the processing was verified graphically, overlapping the binary masks obtained from the softmax network and the output masks obtained after the removal of the small objects (Fig. 16):



(A) (B)

Figure 16 - (A) Original tile; (B) overlapping of the processed mask on the original tile (cores identified in yellow, cores discarded in blue)

2.3.3. Tile Processing

Histopathological tiles were evaluated to eliminate images of reduced quality or containing artifacts due to poor tissue preservation.

Each tile was then scanned for any blurred areas. The presence of artifacts may be due to the different thickness or density of the tissue, as well as the possible presence of any air bubbles.

The extent of the blurring of the tiles was assessed using the Laplacian method. The Laplacian operator performs a second-order derivation, aimed at identifying the points of rapid transition within an image.

Once the transitions (edges) have been identified, the variance of the Laplacian (σ^2) is calculated to measure the dispersion of the image values:

$$\sigma^2 = Var(\nabla^2 I)$$

where $\nabla^2 I$ is the Laplacian operator applied to the image:

$$\nabla^2 f(\mathrm{x},\mathrm{y}) = \frac{\partial^2 \mathrm{f}(\mathrm{x},\mathrm{y})}{\partial \mathrm{x}^2} + \frac{\partial^2 \mathrm{f}(\mathrm{x},\mathrm{y})}{\partial \mathrm{y}^2}$$

Images for which the variance is less than the set threshold value (th=100) are removed from the dataset.

Images that are not blurry are further analyzed to evaluate the percentage of pixels belonging to the slide, to maintain only the tiles that have a high content of pathological tissue. In this case, ROIs that matched the slide by more than 50% were removed from the dataset. Areas associated with the slide can be recognized by setting a threshold on color intensity (th=244/255), which distinguishes white areas from areas of tissue stained in H&E.

The implemented algorithm is represented in the following flowchart:

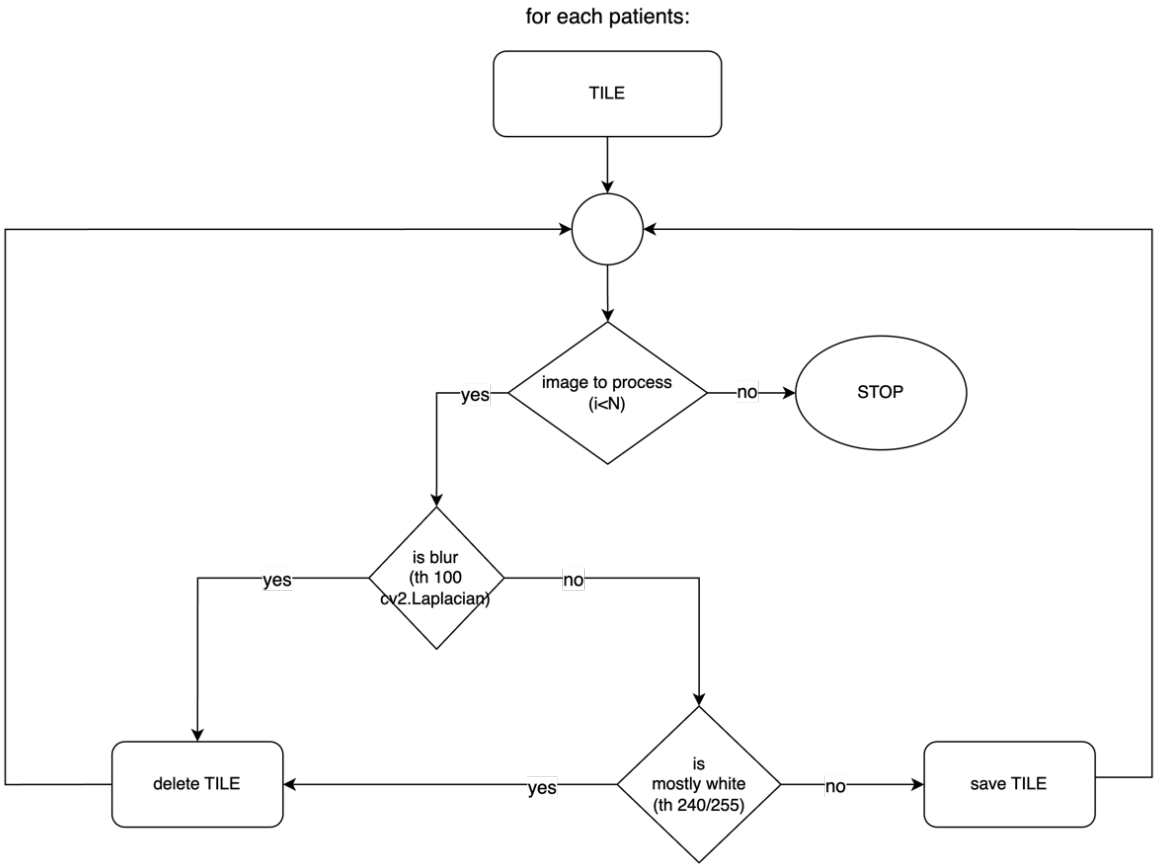


Figure 17 - Tile preprocessing flowchart

The result obtained from the application of the previous method is shown in figure 18.

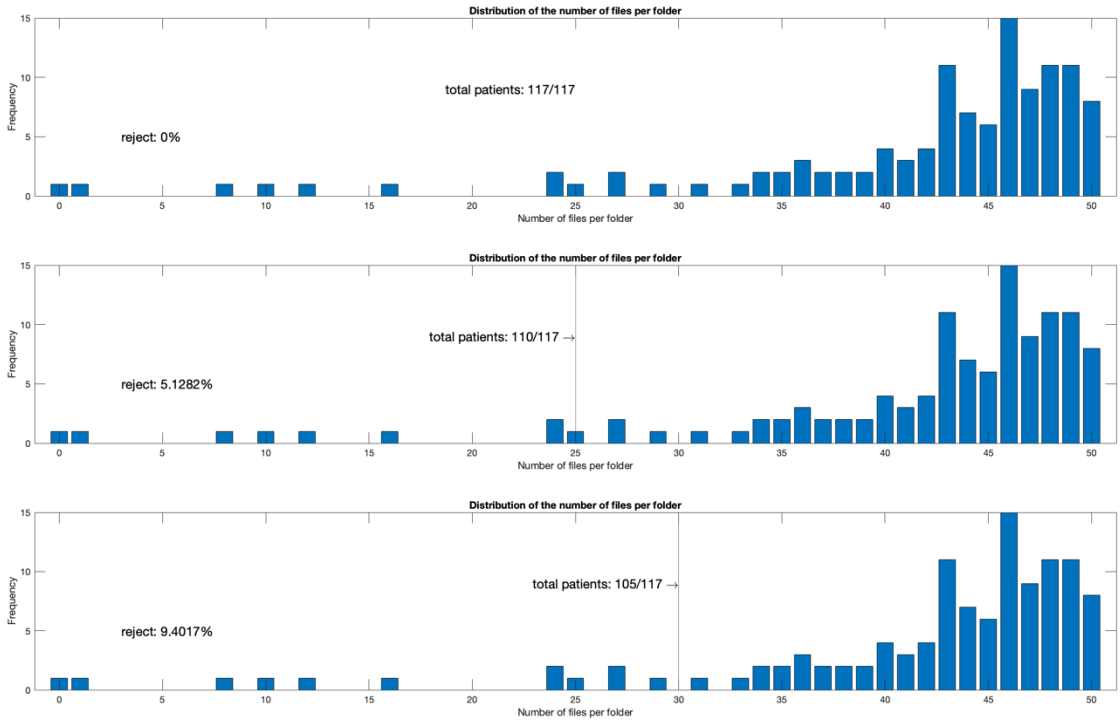


Figure 18 - Result obtained from the processing of the tiles

The image above (Fig. 18) shows the distribution of files contained within each patient file after the cleaning algorithm was applied (Fig. 17), compared to the maximum number of files initially contained in the folder.

Initially, 50 40x histological ROIs were extracted for each patient, with a size of 1024x1024. Random extraction allows you to obtain a dataset that associates the same number of ROIs with each patient, without checking the ROI status.

After applying the tile processing method explained above, each subject will be associated with a different number of images, depending on the number of tiles that have passed the applied quality checks.

Depending on the distribution of the files and therefore the overall content of the data that can be analyzed for the various subjects, three thresholds are evaluated, depending on the minimum number of files acceptable within a medical record to keep the subject valid.

The thresholds evaluated are:

- minimum number of acceptable files equal to 0 files: all patients are included in the study, for a total of 117/117 subjects, with a difference of 0%;
- minimum number of acceptable files equal to 25 files: all patients who meet the chosen inclusion criterion are included, for a total of 110/117 subjects, with a difference of 5.128%;
- minimum number of acceptable files equal to 30 files: All patients who meet the chosen inclusion criterion are included, for a total of 105/117 subjects, with a difference of 9.501%.

Prior to the definition of the final data set, a further image processing phase was applied before proceeding with the final rejection of the tile.

The algorithm in this case provides, after the first identification of the blurred images, by means of the Laplacian method, the manipulation of the HSV color space to improve the saturation and contrast of the original image.

In particular, saturation has increased by 50 units per pixel and contrast by a factor of 1.5.

Increasing these values simultaneously leads to an overall increase in the vibrancy of the image, possibly leading to an increase in the number of files that could be saved within the final dataset, thanks to an improvement in the color of the image.

Images with the most vibrancy are further tested to check the blur status using the previous method.

If the image is sharp this time, the white will be checked before saving or permanently deleting the tile.

The algorithm used is explained in the following flowchart:

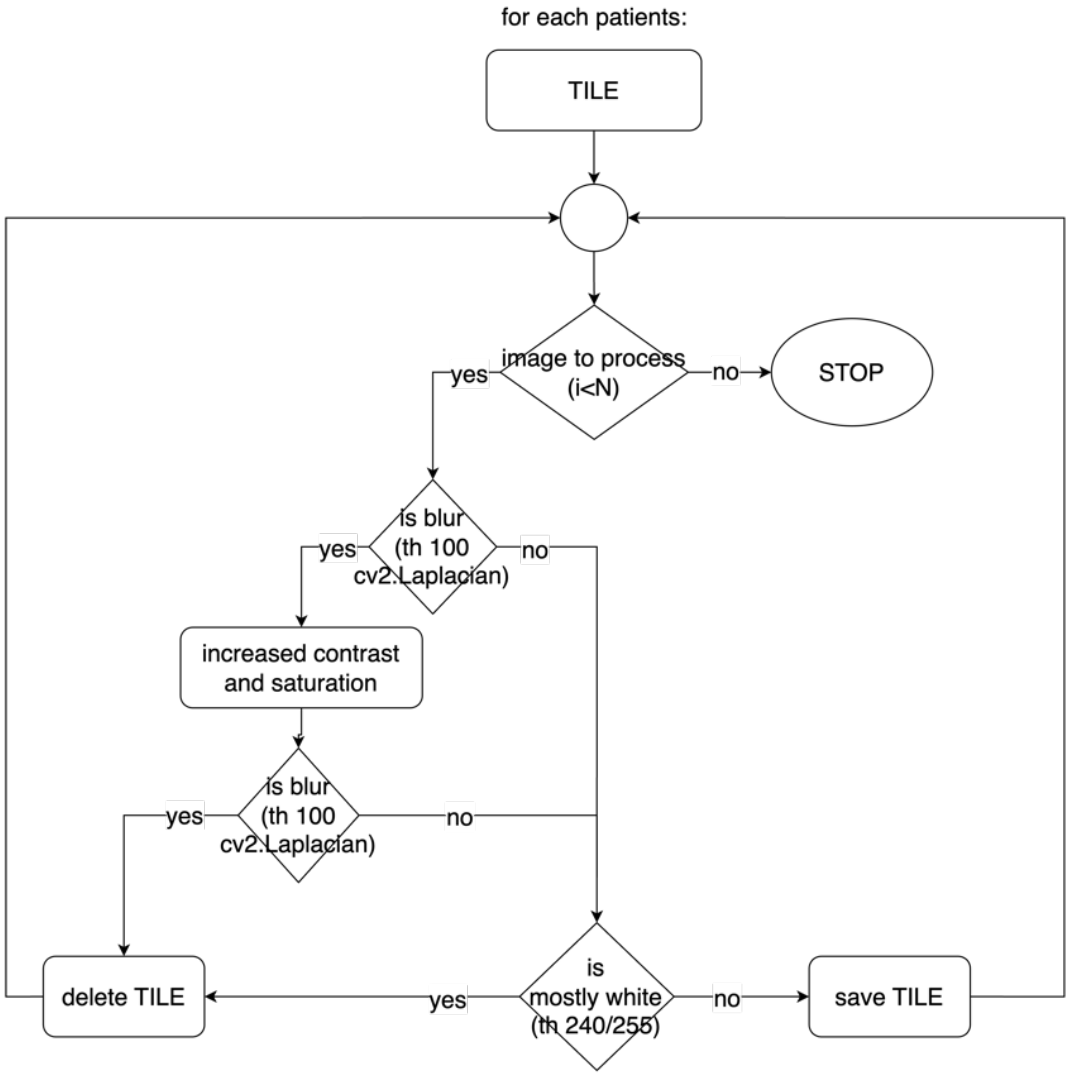


Figure 19 - Optimized tile pre processing

The results obtained after the application of the improvement algorithm (fig.18) are shown in the following figure:

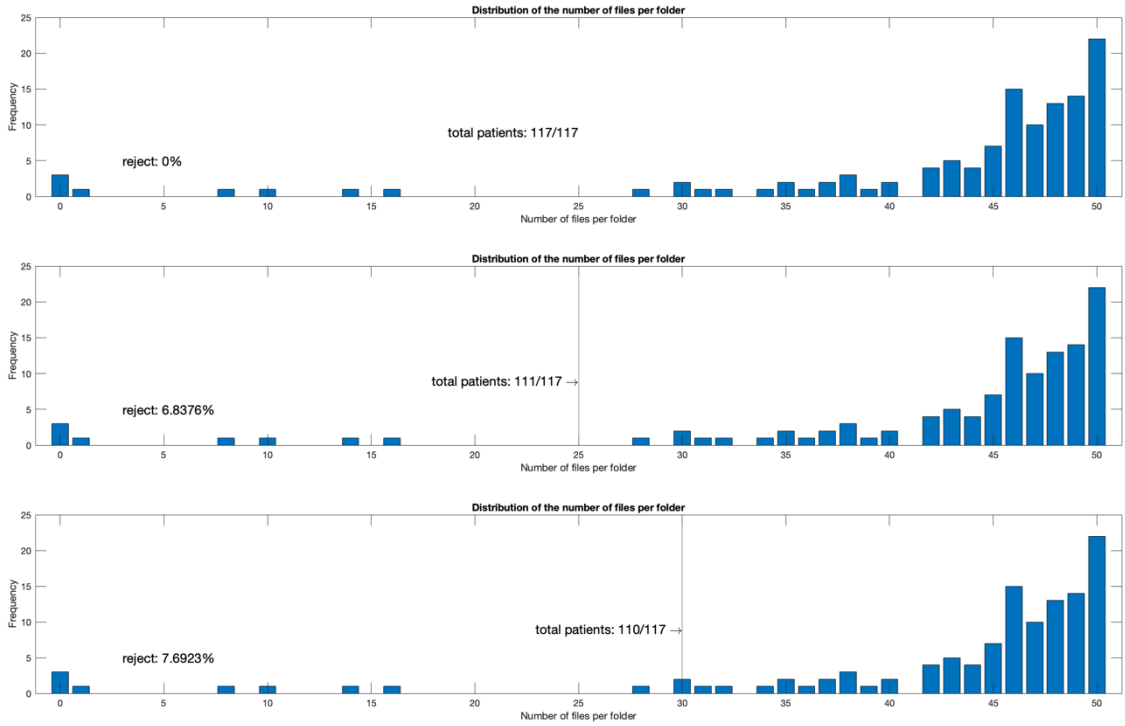


Figure 20 - Result of the application of optimized tile processing

The second method used demonstrates an overall increase in the number of images saved within each patient file. The thresholds evaluated are the same as in the previous case, but with variation in the results obtained:

- minimum number of acceptable files equal to 0 files: all patients are included in the study, for a total of 117/117 subjects, with a difference of 0%;
- minimum number of acceptable files equal to 25 files: all patients who meet the chosen inclusion criterion are included, for a total of 111/117 subjects, with a difference of 5.128%;
- minimum number of acceptable files equal to 30 files: all patients who meet the chosen inclusion criterion are included, for a total of 110/117 subjects, with a difference of 7.692%.

The difference in color and quality between the discarded and saved tiles, obtained after applying the patient inclusion criteria, is shown below:

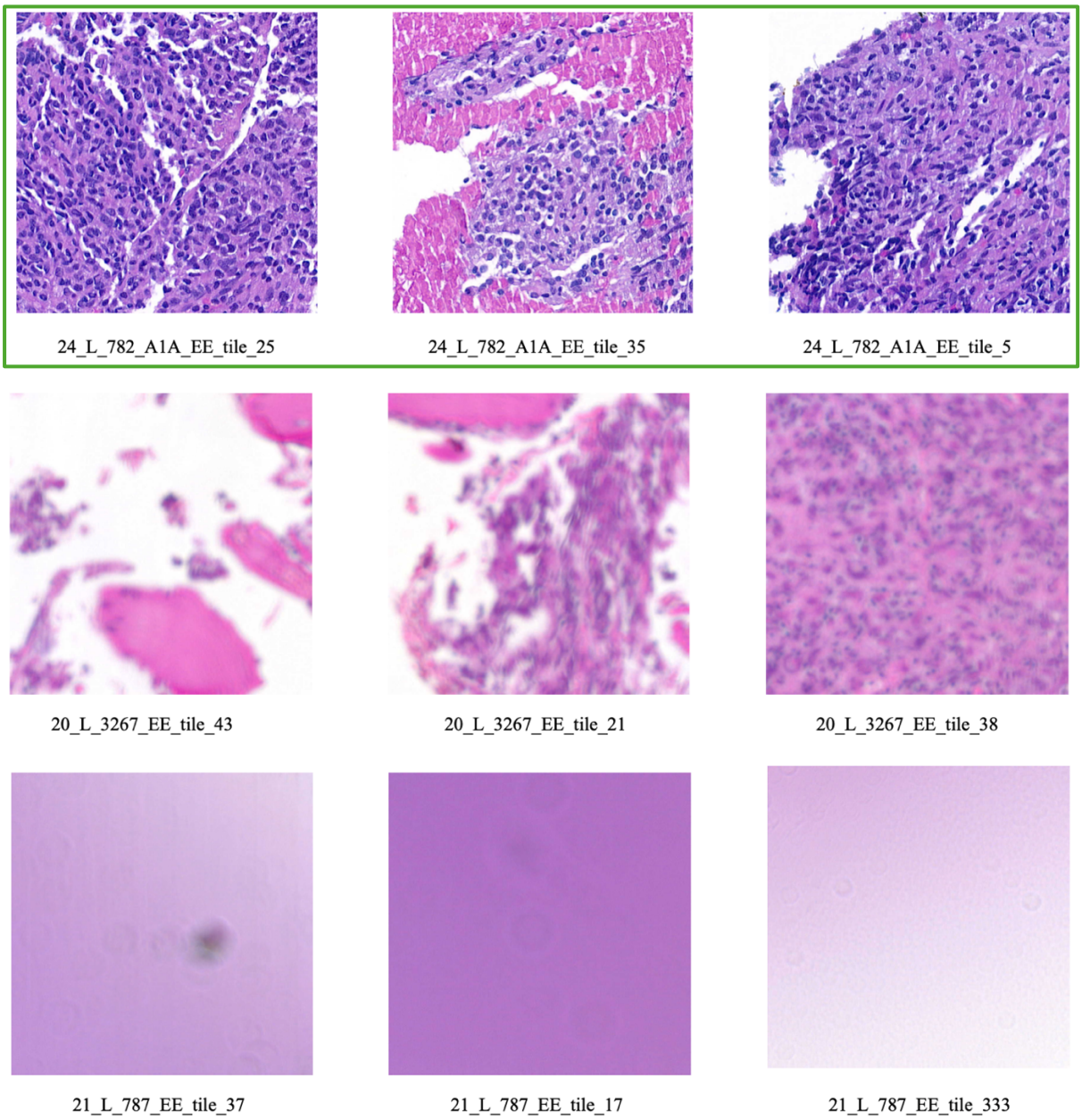


Figure 21- In the green box the optimized and saved tiles; The other images are examples of discarded tiles. The cards refer to different groups of patients, specified in the figure

2.4 Feature extraction

Once the reduced dataset was obtained, different morphological features were extracted from the tiles and masks, so that they were associated with each patient and uniquely characterized him.

The extraction of the features was carried out manually to optimize the characteristics that could be significant for the separation of the subjects into homogeneous groups. Subjects affected by DLBCL, in fact, are highly homogeneous and

poorly separable using only the morphological characteristics extracted from histopathological slides.

Separate analyses were performed on masks and tiles on three regions of interest: the whole image, the small cell population, corresponding to lymphocytes, and the large cell population, corresponding to DLBCL tumor nuclei.

Masks were used to extract 12 geometric features, while tiles were used to extract 6 texture features, of which 4 were extracted on tumor cells and 2 on the global image. In total, 18 morphological features are associated with each patient.

First, the pixel diameter of a DLBCL cell was manually selected, used as a reference to distinguish cancer cells from non-cancer cells (lymphocytes).

Figure 22 shows the manual selection method, which consists of identifying the left and right extremes of the core diameter:

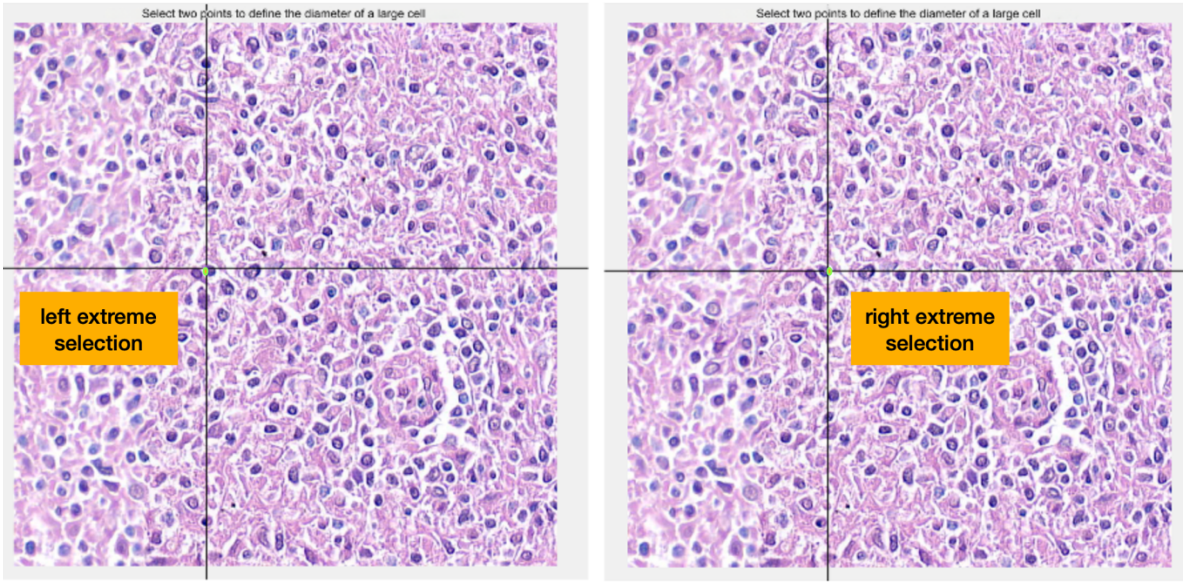


Figure 22 - Manual selection of DLBCL cell diameter extremes: left end selection, right end selection.

The procedure was manualized in the absence of indications or annotations on the histological images analyzed, based solely on chromatic information. The diameter identified is 14 pixels.

However, it is known in the literature that the size of neoplastic B cells reaches dimensions up to 4-5 times larger than healthy cells. [22]

The diameter was then chosen on one of the tumor B cells that could be included in this clinical definition, evaluating the result obtained through the graphical representation of the two cell populations (Fig. 23)

The separation of the two populations allows the extraction of targeted characteristics to extract, separately, the indicators that characterize lymphocytes from the indicators that characterize DLBCL cells. The extracted parameters will be analyzed later.
The two identified cell populations are shown in the figure below:

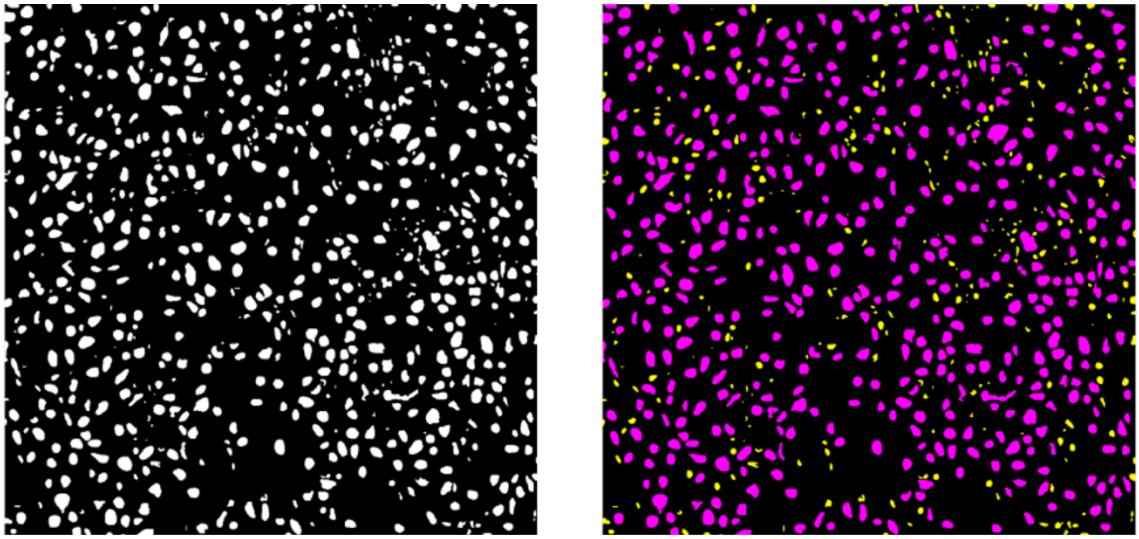


Figure 23 - (A) mask of the nuclei; (B) identified cell populations: in yellow the lymphocytes, in purple the DLBCL cells

Based on the available data, both shape and texture features were extracted to obtain a set of morphological characteristics associated with each patient that represented on average the characteristic cell population of the tumor. The matrix of average characteristics was finally normalized by min max normalization, obtaining variables

between 0 and 1, so that each had the same weight once entered the clustering algorithm:

$$x_{i,norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Below is the operational flow diagram for feature extraction, used for subject clustering:

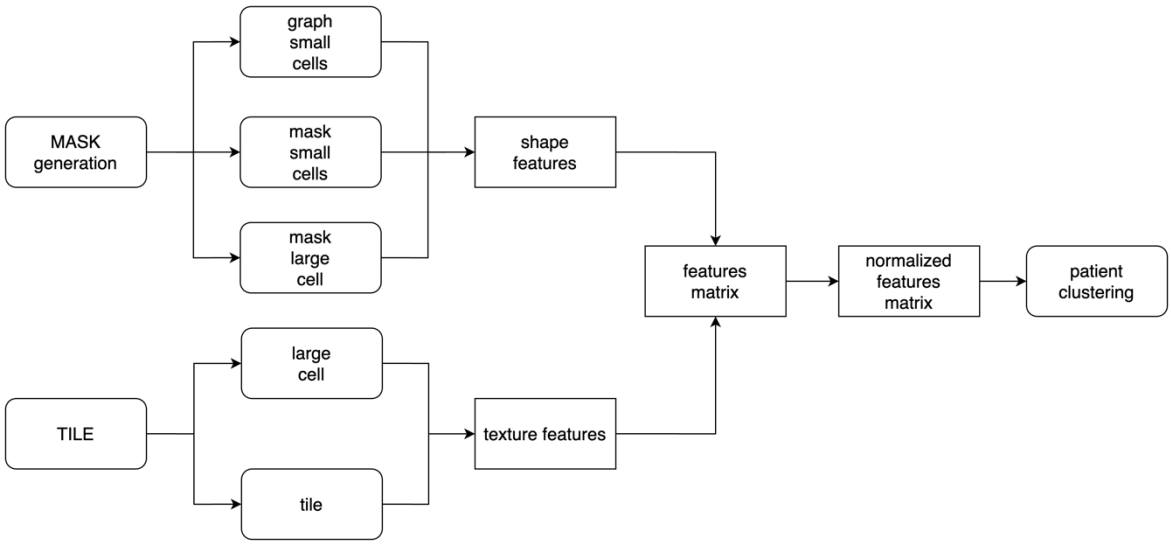


Figure 24 - Feature extraction flowchart

2.4.1 Features extracted from binary masks

On the masks, 12 geometric features were extracted, respectively extracted from the small cell graphs, the small cell population, the large cell population and the overall mask.

The extracted features are shown in the flowchart below.

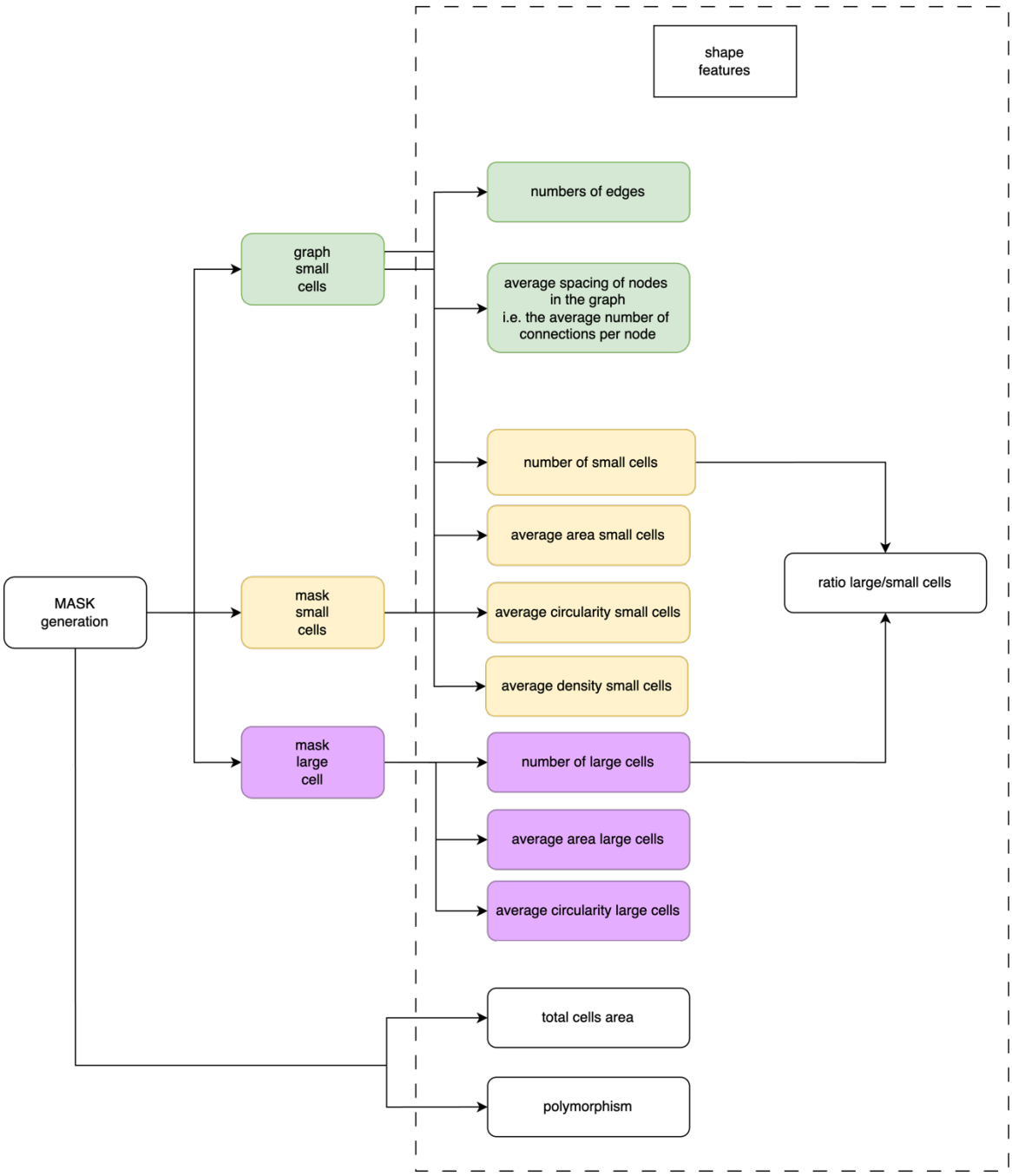


Figure 25 - Flowchart of features extracted from binary masks of nuclei

The following are the characteristics extracted from the two cell populations identified by diameter threshold:

- total number of cells belonging to the cell population;
- average area of the cell population;
- average circularity of the cell population.

In the case of lymphocytes, the average density of small cells was also extracted.

On the total number of cells belonging to the various populations, the ratio was derived, as the number of large cells compared to the number of small cells and the sum of the two cell populations (total cells area).

The identified ratio emphasizes the subjects that present a disproportion between the two populations:

$$R = \frac{\sum large\ cell}{\sum small\ cell}$$

with R < 1 when the lymphocyte population is greater than the number of DLBCL tumor cells present in the tiles.

The values of the total cells area, the number of small cells, and the number of large cells were divided by the total number of tiles for each subject. This is because, after the application of the pre-processing operations, the subjects present a different number N of tiles, caused by the removal of the tiles that have not passed the quality controls carried out.

Finally, the polymorphism parameter, a measure of cell size variability, was obtained on the masks, as the standard deviation of the cell areas identified in the segmentation.

2.4.1.1 Graphs

A graph is a mathematical representation useful for extracting morphological features on histological tissues.

The matlab function used to create the graph is: '$G = graph(s, t)$'.

The unoriented graph returned by the function consists of a series of nodes (or vertices) and a set of edges, which connect the nodes, without a specific direction of connection. In tissue image analysis, the nodes are represented by the centroids of the segmented nuclei, while the edges represent the connections between the centroids themselves.

The use of cell graphs in the study of oncological pathologies is significant because it allows to extract the spatial arrangement of cells and their interactions, allowing to capture complex details on the architecture of tissues. In histological image analysis, graphs can model both low-level features (such as cell shape or size) and high-level features (such as tissue structure).

Figure 26 shows on the left the centroids superimposed on the segmented nuclei, and on the right the graph made on all the cellular areas identified in the mask.

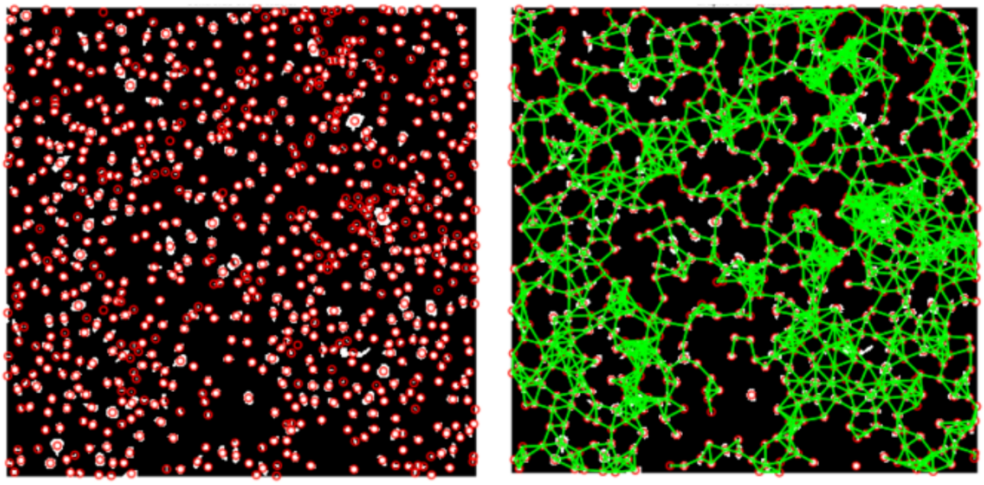


Figure 26 - Superposition of centroids (A) and graph (B) on the binary mask of nuclei

The extracted features were calculated on the graph built on the population of small cells, related to lymphocytes, as shown in figure 27.

Figure 27 - Example of a graph obtained from the population of small cells

The characteristics extracted from the lymphocyte population are:

- number of nodes (corresponding to the total number of lymphocytes);
- number of arcs;
- Average node spacing in graphs (average number of connections per node).

In the analysis of histological images, the use of the graph and the characteristics extracted from it can be useful for the representation of cell populations and for the relationships between them.

It was decided to limit the characteristics extracted on the graph to the lymphocyte population only, starting from the hypothesis that a region with a high presence of benign

B cells would be associated with a better outcome of DLBCL pathology, since it would trigger favorable immune mechanisms. This hypothesis was provided directly by the referring pathologist.

2.4.2 Features extracted from histological tiles

Texture features were extracted from the histological tiles both on the entire ROI region and on the areas corresponding to DLBCL tumor cells.

Parameters derived from the gray levels co-occurrence matrix (GLCM) were calculated on tumor cells. The GLCM matrix describes the texture of an image by measuring how often the pixel pairs have specific values, relative to a specific spatial relationship of pixels that occurs in an image. [16]

In particular, the matrix was calculated in *Matlab* using the '*graycomatrix' function*, with angle 0 (*Offset = [0 1]*). Among the possible directions, shown in figure 28, the GLCM was calculated based on the horizontal variation of the pixels (one pixel away to the right).

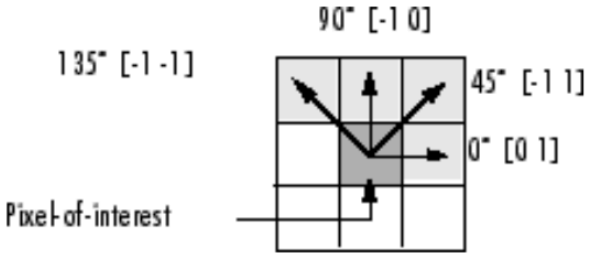


Figure 28 - Possible offset directions [23]

The texture characteristics extracted in the study, obtained from DLBCL cells using the '*graycoprops'* function, are:

- Contrast: Calculated as the difference in gray between adjacent pixels, such as: [23]

$$Contrast = \sum_{i,j} |i-j|^2 P(i,j)$$

  where $P(i,j)$ represents the probability of distribution of the difference in gray level between adjacent pixels (i,j). Contrast indicates the amount of local variation in the image, where a high value corresponds to a highly heterogeneous image.

- Correlation: Measures the degree of similarity between adjacent pixels, for example: [23]

$$Correlation = \frac{\sum_{i,j}(i-\mu_x)\left(j-\mu_y\right)\cdot\ P(i,j)}{\sigma_x\sigma_y}$$

where: μ_x, μ_y represents the average of the gray levels of the row and column, respectively; σ_x, σ_y the variance of the gray levels of the row and column.

- Energy: Also known as, angular momentum second, represents a measure of the uniformity of the GLCM, as the sum of the squared elements in the GLCM:[23]

$$Energy = \sum_{i,j} P(i,j)^2$$

- homogeneity: measures the proximity of the distribution of elements in the GLCM to the diagonal, such as: [23]

$$Homogenity = \sum_{i,j} \frac{P(i,j)}{1+|i-j|}$$

Below, in figure 29, is shown the flowchart that schematizes the functions extracted on the tiles.

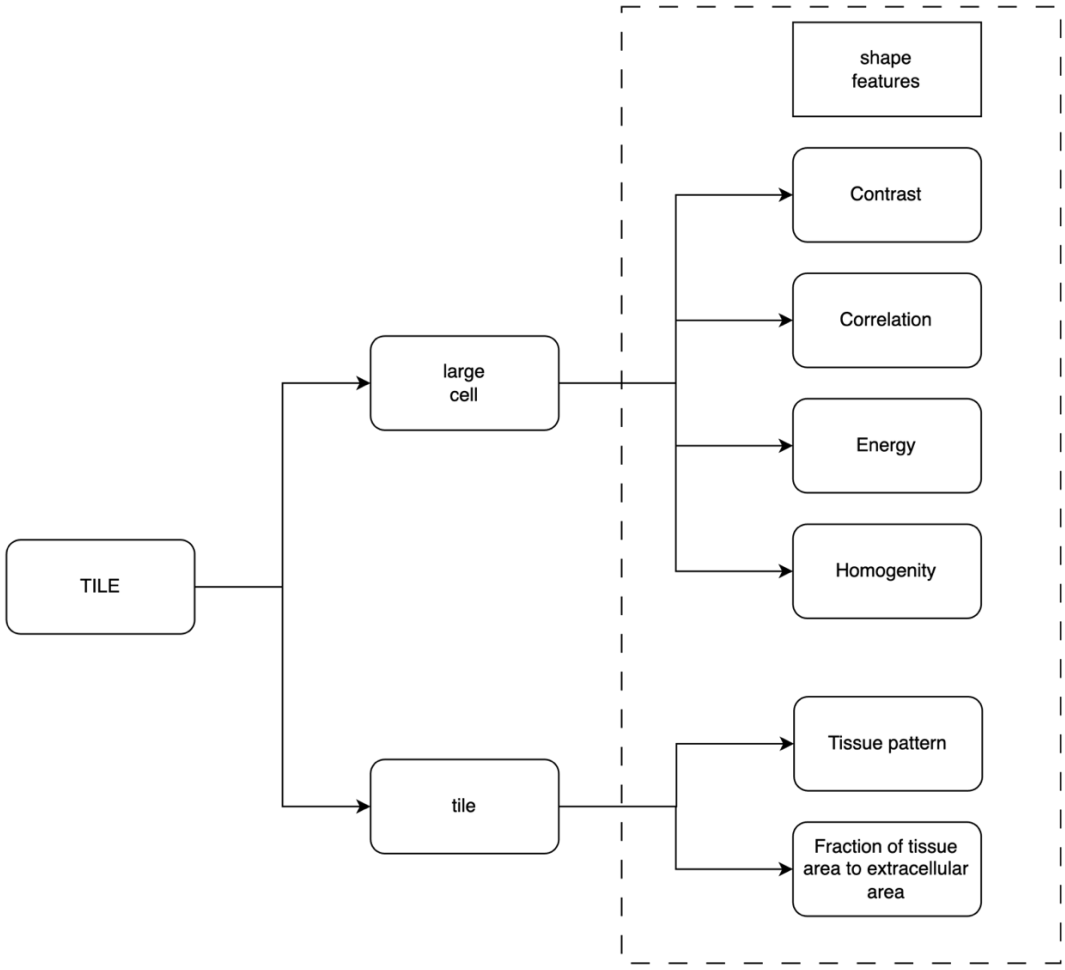
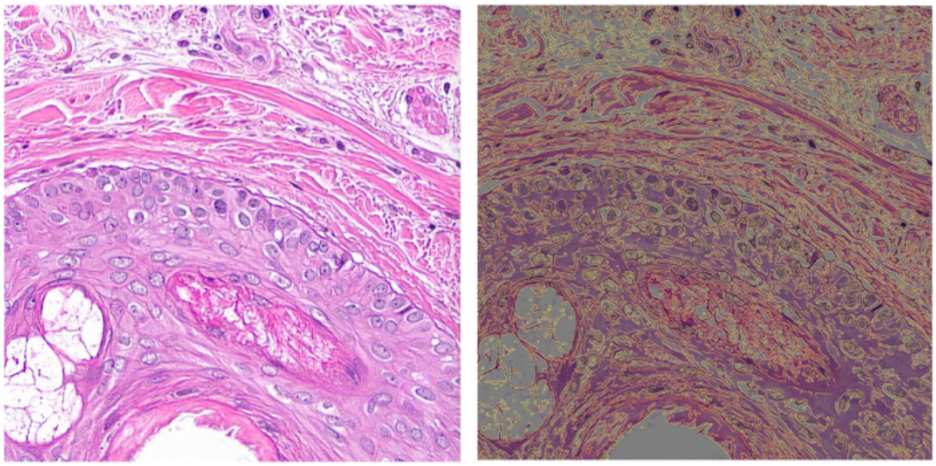


Figure 29 - Flowchart of the extraction of texture features extracted from histopathological tiles

Two further parameters were obtained on the tiles: the tissue density and the fraction of tissue area on the extracellular area.

To identify the tissue pattern, the *'edge(gray_img, 'Canny')*' function was used, applied to the grayscale image, which is useful for detecting edges in the image. Once the number of pixels classified as "edges" was obtained, the tissue pattern was calculated as the number of pixels associated with the edges divided by the totality of pixels in the image. The identified model is shown in the following figure:



(A) (B)

Figure 30 - (A) histological tile; (B) the edges identified in tile A are highlighted in yellow

Finally, to identify the fraction of tissue area on the extracellular area, a binary segmentation algorithm based on color threshold was used to identify the two portions of interest: the area relative to the tissue and the area relative to the extracellular matrix. From a morphological point of view, this parameter can provide information on tumor architecture, useful for differentiating patients.

2.5 Patient clustering and survival analysis

The normalized feature matrix was tested on two clustering algorithms: K-means and hierarchical clustering. Each method was evaluated to obtain k=2,3,4 clusters of patients.

To evaluate the quality of the clusters obtained, the '*silhouette'* method was used, in particular:

- Silhouette mean: calculated average value of the silhouette over all points of the dataset, it represents a global measure of the quality of clustering. The silhouette value for a point *i* is calculated as [17]:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ represents the average distance between point *i* and all other points within its cluster and $b(i)$ the average distance between the point i and all points in the nearest cluster.

$S(i)$ values have range from -1 to 1, where 1 represents an optimal that the point is well associated with one's cluster and poorly matched to the other clusters; Null or negative values indicate that the split may not be adequate.

The average value $\overline{S}(i)$ represents the average of all the values of the silhouette.

- Silhouette plot: This is a graphical method of displaying values $S(i)$ for each point in the cluster. In the chart, each bar represents a point, and each point is grouped by cluster. The distribution of each group is represented graphically by the Euclidean method of distance.

The clusters obtained were visualized by reducing the dimensionality of the variables to two main components, through the analysis of the main components (PCA). This technique was used to transform normalized variables into a set of uncorrelated variables, known as principal components. Variables can thus be displayed in a new space with reduced complexity, while maintaining their informative content. The first principal component is the direction along which the data varies the most, the second principal component is the direction orthogonal to the first that has the second greatest variance, and so on.

In the graphs, observations that belong to the same cluster (depending on the result of the clustering) will be colored in the same way, allowing you to see how the data is distributed and how far it is separated from each other.

In the available data, each patient is associated with two main events:

- death: binary vector that represents the state of death of the subject (death=0, death event that did not occur, death=1, death event that occurred);
- any firt event: binary vector that represents the occurrence, in the subject, of a first adverse event (AFE=0, adverse event that did not occur or censored, AFE=1,

adverse event that occurred, or not censored). Aggravations or recurrences of the disease are considered an adverse event;

- date of any first event: represents the time, in months, in which the first adverse event occurred. For subjects in whom AFE did not occur, the time is equal to NaN.

These data, reported for example in Table 1, were used to derive the survival curves of the groups of subjects.

For each clustering, the recurrence-free or worsening survival curve of the population groups obtained using the matlab function '*ecdf*', with the '*function'*, '*survivor'* option, was estimated. This function, used in '*survivor'* mode, returns the survival curve of a group of subjects, like the Kaplan-Meier method. The function returns the empirical cumulative distribution, which represents the probability that the adverse event has occurred up to a certain point in time.

If a curve remains high (survival = 1) for an extended period of time, it means that most patients remain free from long-term recurrence or worsening, suggesting a positive treatment or a better prognosis for that cohort. Conversely, if a curve descends rapidly, it means that many patients are experiencing recurrences or worsening in a short period of time, suggesting a worse prognosis.

Finally, the probability curves obtained were associated with the state of death of the subject himself, to verify the correlation between the two events examined.

Table 1 – Major events associated with the patient: death status, any first event, time any first event

ID	Death (0: No, 1 Yes)	AFE (0: No, 1 Yes)	Time AFE (months)
19_L_2084_A1_EE	0	1	25
19_L_2934_4_EE	1	1	22
19_L_2934_8_EE	0	1	17
20_L_1001_A1_EE	1	1	18
20_L_1001_EE	0	0	Nan
20_L_1020_A1_EE	0	0	Nan
…	…	…	…

Results

3.1 Analysis of the population of subjects

The 111 subjects analyzed were graphically represented as a function of time in months in which the first adverse event occurs, or not.

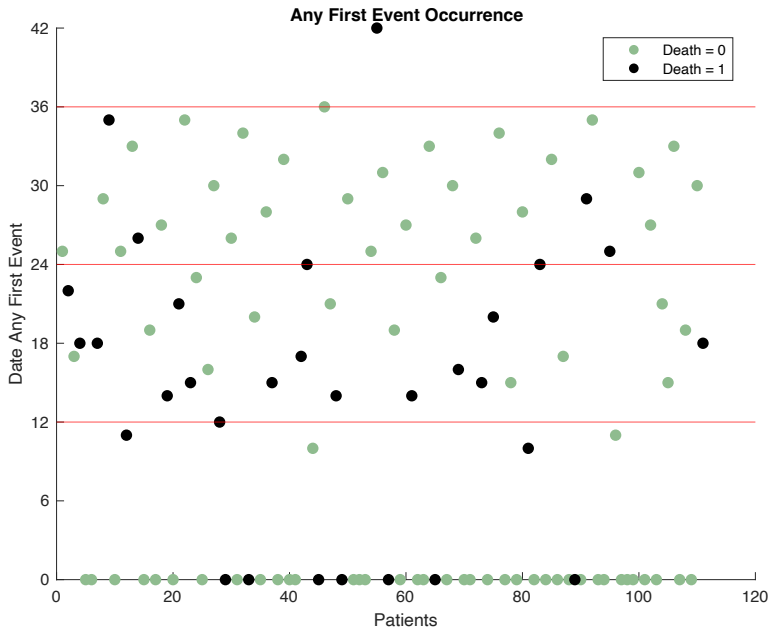


Figure 31 - Occurrence in months since the occurrence of the first adverse event

On the x-axis, at time zero, all subjects who have not had an adverse event are represented. In the subgroup considered, 7 subjects died (in black).
According to the various time bands, the subjects who had an adverse event were then represented. Within the first two years of diagnosis, most subjects died, while most subjects survive in patients who present with the first adverse event two years after diagnosis. The observation period of subjects ranges from up to three years, with only one subject monitored at 42 months.

It is not known what factor leads to the shown result. The increase in deaths in the first 12 months could be due to a greater aggressiveness of the disease, the presence of any external risk factors, the presence of comorbidities or the ineffectiveness of the therapeutic treatment followed.

3.2 Comparison of clustering algorithms

The results of clustering were compared according to the number of subgroups into which patients were divided, respectively with the K-means algorithm and with the hierarchical clustering algorithm (method *'ward'*).

For each algorithm the following are shown:

- the graphical representation of the division into clusters by displaying the first two main components obtained from the PCA;
- silhouette plot for the visualization of the average silhouette of each subject, for the visualization of the goodness of the clusters obtained;
- survival curve for adverse events, with a graphical representation of proportional deaths: the survival curve obtained for each cluster of subjects is superimposed on circles that represent, in proportion to their size, the number of deaths that occurred at that moment in time.

For clustering obtained by means of a hierarchical division algorithm, the dendrogram of the output obtained on the normalized variables is also provided, with visualization of the cut and the consequent clusters formed.

3.2.1 Splitting into two clusters

This section compares the results of splitting into two clusters.

- Clustering K-means:

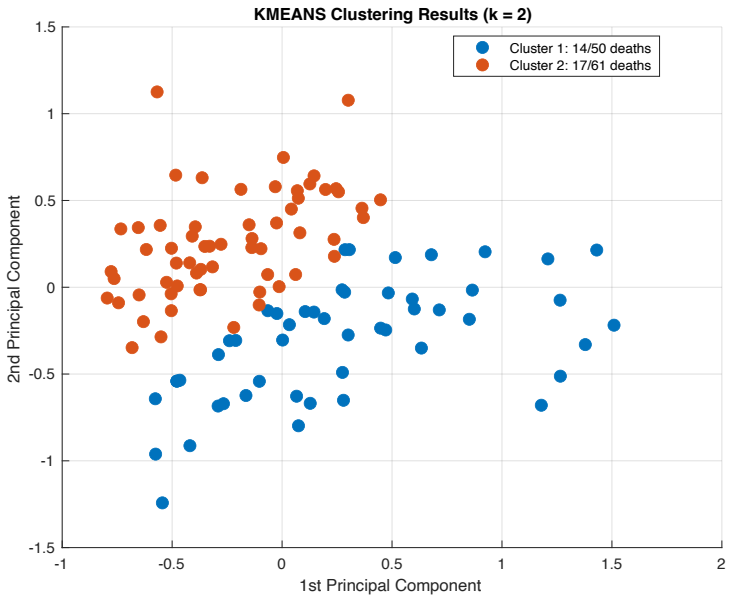


Figure 32 - The result of the grouping obtained with the k-means algorithm (k=2), with the number of deaths, displayed by the first two main components.

Figure 32 shows the two groups obtained using the K-means algorithm, for k=2, on normalized features. The result is represented graphically as a function of the first and second main components obtained by the PCA.

The two clusters, although numerically different, have roughly the same mortality rate (~28%).

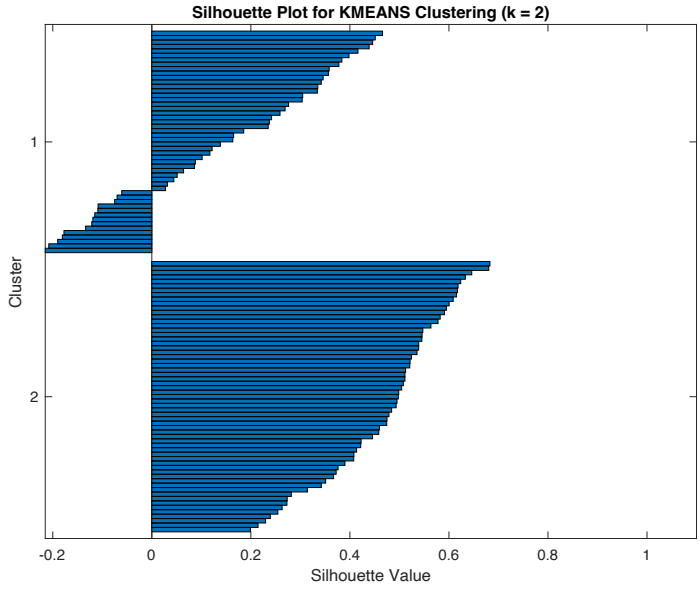


Figure 33 - Silhouette graph for the k-means algorithm (k=2)

The Silhouette graph shows, in cluster 1, 19 subjects, of which 5 with a silhouette value less than or equal to 0.1 and 14 with a negative value.

This shows an equidistance from both clusters for subjects with $S(i) \leq 0.1$ and a probable misassignment to the cluster for subjects with $S(i) < 0$

Most of the subjects belonging to the clusters have $S(i) \geq 0.2$.

Cluster 2 has higher silhouette values overall than cluster 1.

The value of the mean of the silhouette is equal to $\overline{S}(i) = 0.3221$.

Figure 34 shows the adverse event survival curves obtained in the two clusters:

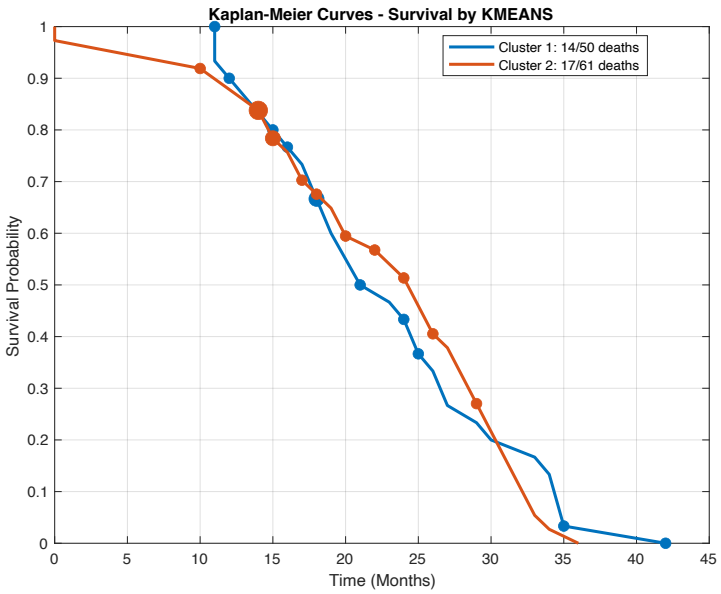


Figure 34 - Adverse event survival curve for the k-means algorithm (k=2)

From the graph, the curves are hardly distinguishable, with a trend reversal around 17 months. From the division obtained with the K-means method (for k = 2) the two groups obtained show a similar evolution over time with respect to the variable of interest (adverse event). The two groups are therefore not significantly separable using the morphological features extracted from histopathological images. As a result, the clusters obtained show no difference in survival.

- Hierarchical clustering:

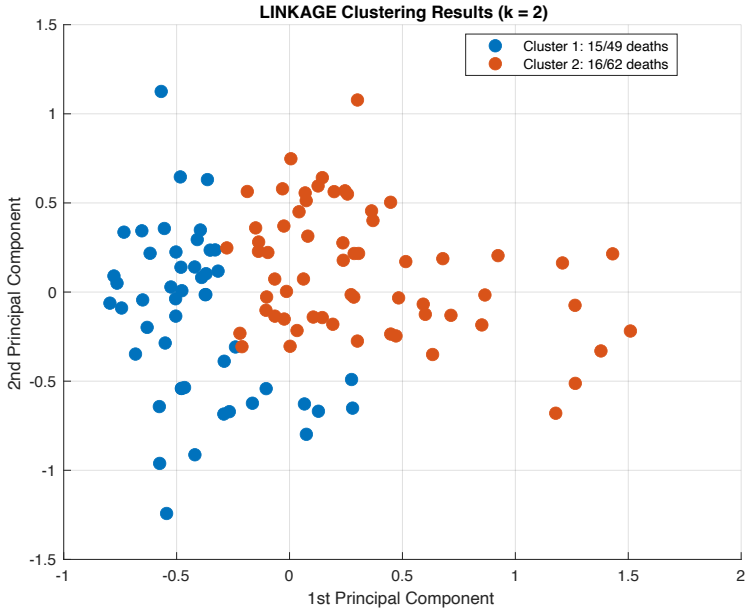


Figure 35 - Result of the grouping obtained with the linkage algorithm (k=2), with number of deaths, displayed by means of the first two main components.

Figure 35 shows the two groups obtained from the use of the hierarchical clustering algorithm '*linkage'*, for k=2, on the normalized characteristics.

Also in this case, cluster 2, in red, is numerically more populated. The mortality rate in the two groups is different, with 30.61% associated with group 1 and 25.81% for group 2.

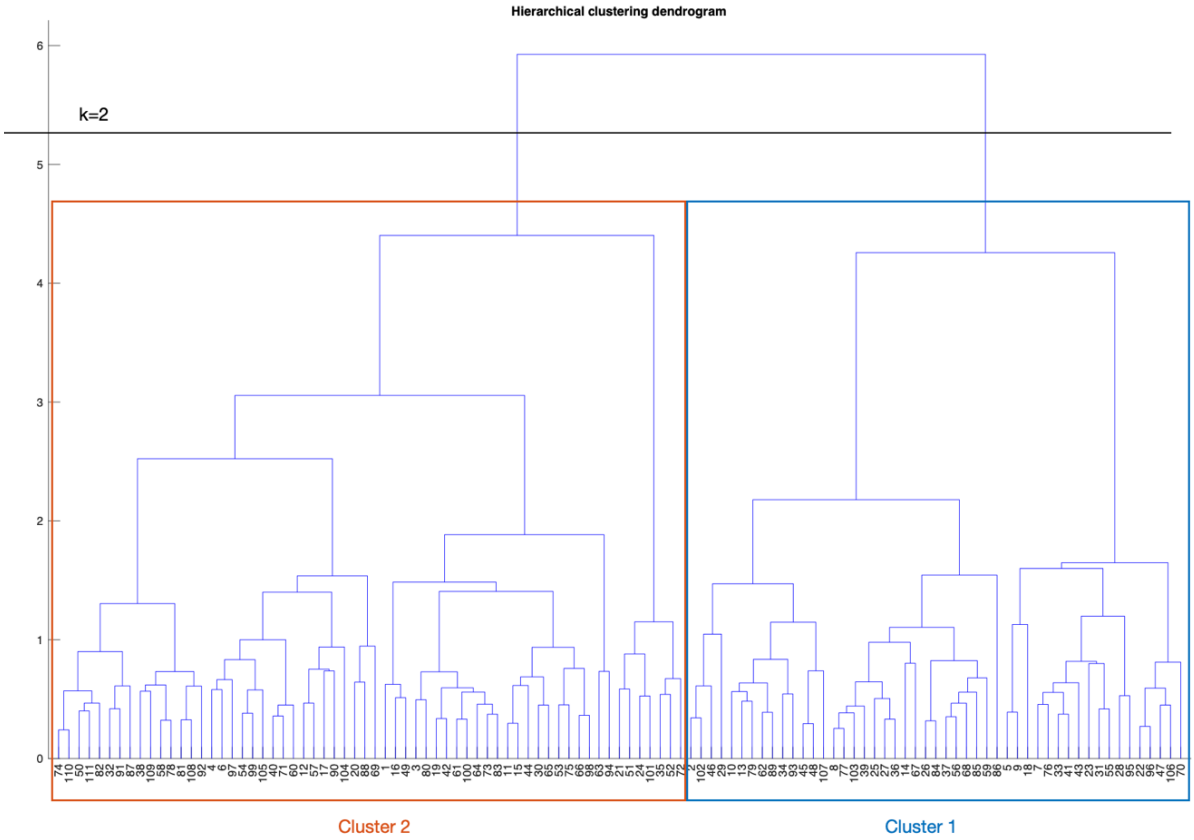


Figure 36 - Dendrogram representative of hierarchical clustering obtained from normalized features (k=2)

The dendrogram represents in figure 36 the division of patients into two clusters.

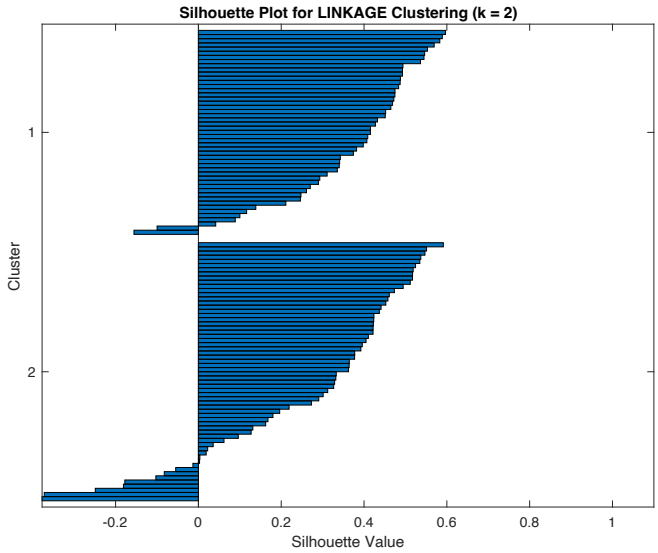


Figure 37- Silhouette graph for the linking algorithm (k=2)

The Silhouette graph (Fig. 37) shows 11 subjects with a negative value, distributed in both clusters (2 in cluster 1, 9 in cluster 2). The clusters have similar silhouette values, showing similar behavior in terms of internal cohesion and separation between the clusters. The value of the mean of the silhouette is equal to $\overline{S}(i) = 0.3112$.

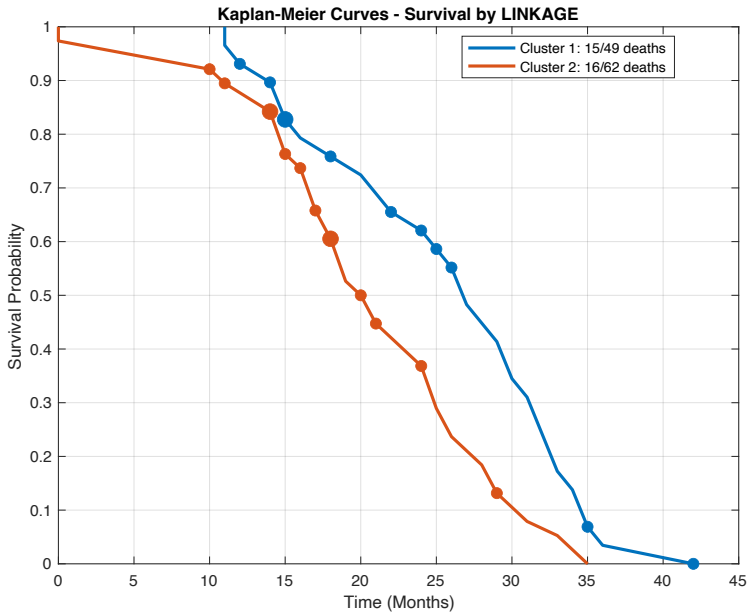


Figure 38 - Result of the grouping obtained with the linkage algorithm (k=2), with number of deaths, displayed by the first two main components

The adverse event survival curve obtained (Fig. 38) is more distinct for the two groups of subjects identified by the hierarchical clustering algorithm. Cluster 1, in blue, has better survival of the event than cluster 2, in red, which has lower values and a greater slope. Around 15 months, the curves partially overlap, indicating some similarity in the likelihood of survival between the two groups at that time. However, in the following moments of time, the curves are widely separated. This divergence between the curves indicates a difference in long-term survival rates between the two groups.

The two groups identified by the hierarchical clustering algorithm lead to the formation of the two populations that have different intrinsic characteristics or risk factors, which influence their probability of surviving over time.

Table 2 summarizes the parameters that characterize the groups identified by the two algorithms tested.

Table 2 - Comparison of the clusters parameters (k=2)

Method	*Clusters*	*Percentage of deaths*
Clustering K-means	Cluster 1	28.00%
	Cluster 2	27.87%
Hierarchical clustering	Cluster 1	30.61%
	Cluster 2	25.81%

3.2.2 Division into three clusters

This section compares the results of splitting it into three clusters.

- Clustering K-means:

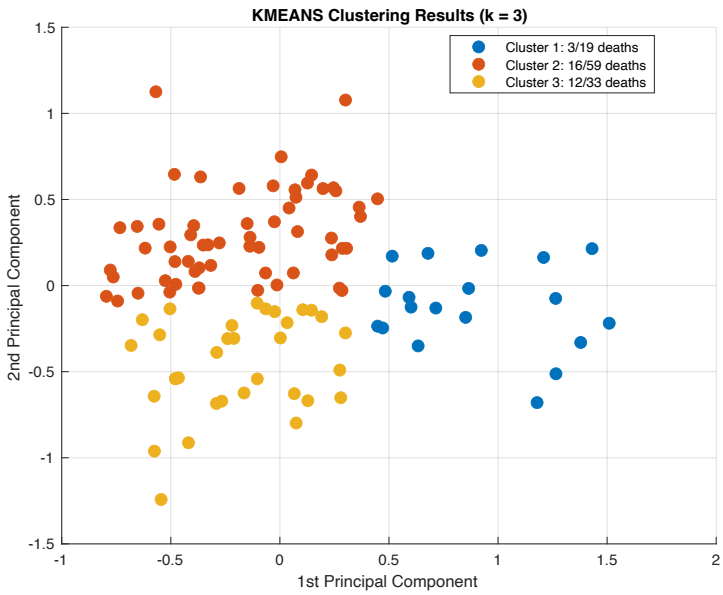


Figure 39 - The result of the grouping obtained with the K-means algorithm (k=3), with the number of deaths, displayed by means of the first two main components.

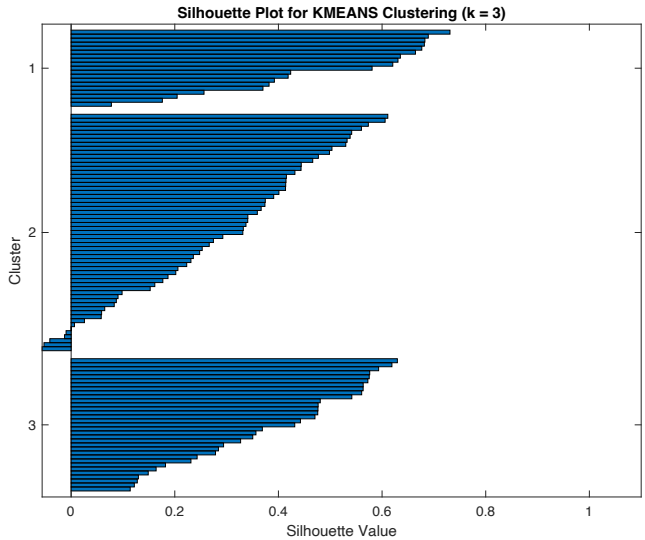


Figure 40 - Silhouette graph for the K-means algorithm (k=3)

In Figure 39 you can see the patient groups created by the K-means algorithm, for k=3. This subdivision, as shown by the graph of the silhouette (Fig. 40), is more accurate than the division into two clusters previously analyzed, carried out in a similar way on the morphological characteristics extracted from the entire population of subjects.

The following are highlight:

- fewer subjects with possible assignment to the wrong cluster: the value has dropped from 19 (k=2) to 5 (k=3);
- Increase in the average value of the silhouette: the average value of the successful silhouette equal to $\overline{S}(i) = 0.3508$.

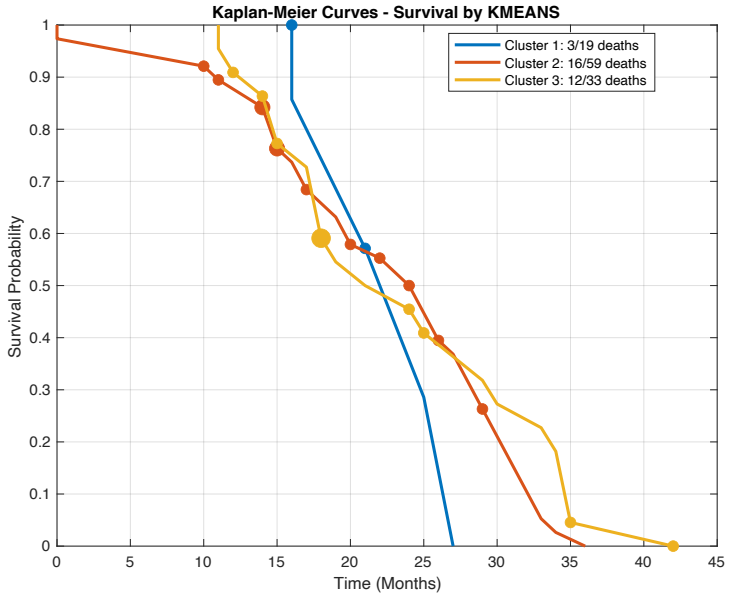


Figure 41 - Adverse event survival curve for the K-means algorithm (k=3)

The trend of the adverse event survival curves (Fig. 41), as in the case of clustering with K-means for k=2, shows curves that tend to overlap for clusters 2 and 3, and a slightly different trend for cluster 1. The latter, in fact, has a higher probability of survival to the adverse event, a trend that is also reflected in a lower effective mortality.

In addition, the curve of cluster 1 falls rapidly after 21 months, showing an opposite trend compared to the other clusters.

Clusters 2 and 3 have approximately the same risk of worsening and recurrence, as shown by the overlapping curves; Cluster 3 is, however, associated with the highest mortality rate (36.36%).

- Hierarchical clustering:

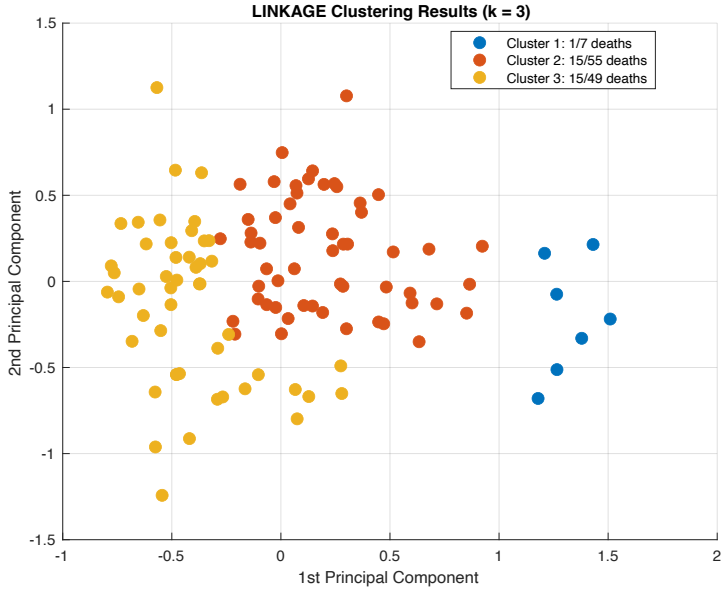


Figure 42 - Result of the grouping obtained with the linkage algorithm (k=3), with number of deaths, displayed by means of the first two main components.

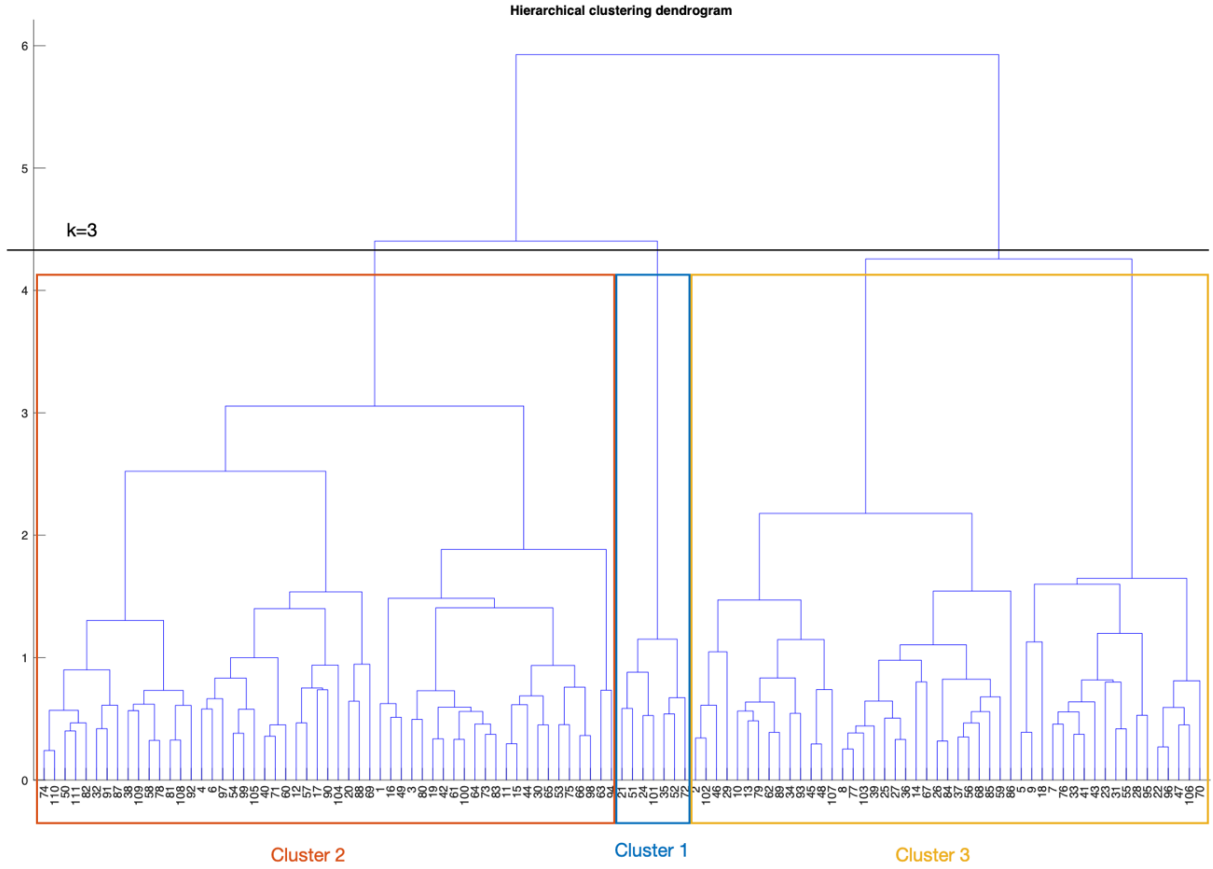


Figure 43 - Dendrogram representative of hierarchical clustering obtained from normalized features (k=3)

The division into three clusters carried out by the hierarchical clustering algorithm (Fig. 42-43) leads to the formation of two groups of patients of the same size (cluster 2, in red and cluster 3, in yellow) and to the formation of a small cluster (cluster 1, in blue), consisting of only seven subjects.

Subjects belonging to cluster 1 show similar morphological characteristics, but very distant from cluster 2, as can be seen from the height of the segment that unites them on a hierarchical scale.

The height of the branches of the dendrogram, in fact, reflects the dissimilarity between the various groups.

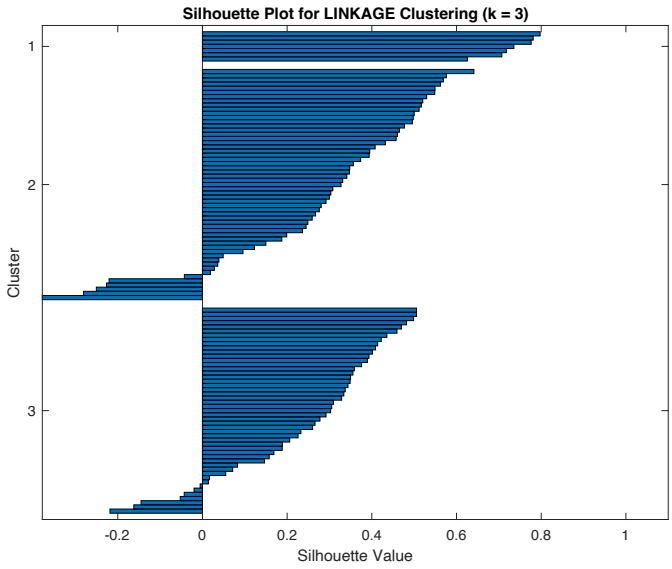


Figure 44 - Silhouette graph for linkage algorithm (k=3)

The silhouette graph (Fig. 44) associated with the grouping shows a high heterogeneity of the characteristics of cluster 1, represented above. The values achieved per silhouette fluctuate, in fact, from a minimum of 0.6 to a maximum of 0.8.

The other two clusters, on the other hand, have silhouette values that tend to be lower, with 6 subjects likely incorrectly clustered in cluster 2 and 7 subjects likely to have incorrect groupings in cluster 3. The average value of the silhouette is equal to $\overline{S}(i) = 0.2942$.

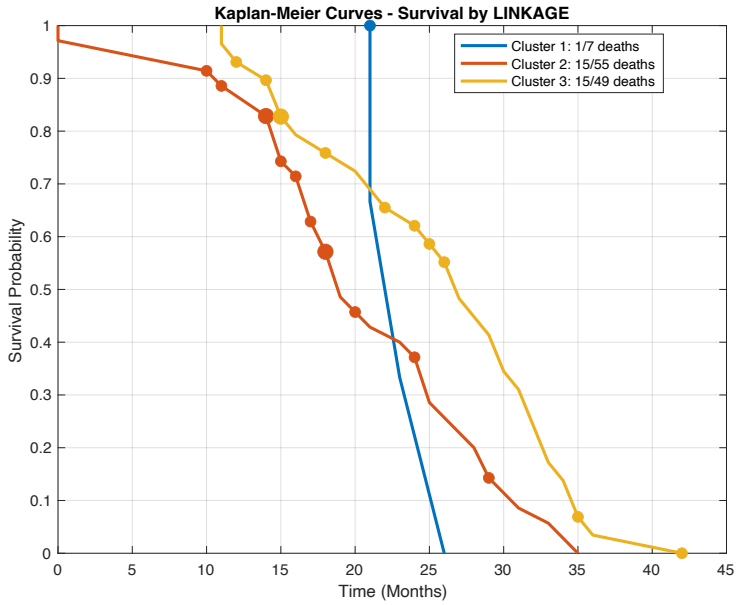


Figure 45 - Result of the grouping obtained with the linkage algorithm (k=3), with number of deaths, displayed by means of the first two main components

The adverse event survival curves (Fig. 45) are quite distinct for clusters 2 and 3, reaching a point of similarity around 15 months.

The curve of cluster 1, on the other hand, is drastically inclined, a trend that is reflected in a better survival of the initial event and a sudden worsening over time. Subjects belonging to cluster 1 present, if not censored, adverse events that occurred after twenty and before twenty-six months.

In contrast to the division obtained with k=2 (Fig.38), the survival curves reflect less the negativity of the prognosis associated with the subjects belonging to the clusters. Table 3 summarizes the parameters that characterize the groups identified by the two algorithms tested.

Table 3- Comparison of the clusters parameters (k=3)

Method	*Clusters*	*Percentage of deaths*
Clustering K-means	Cluster 1	15.79%
	Cluster 2	27.12%
	Cluster 3	36.36%
Hierarchical clustering	Cluster 1	14.29%
	Cluster 2	27.27%
	Cluster 3	30.61%

3.2.3 Division into four clusters

The results of the four-cluster split are compared below.

- Clustering K-means:

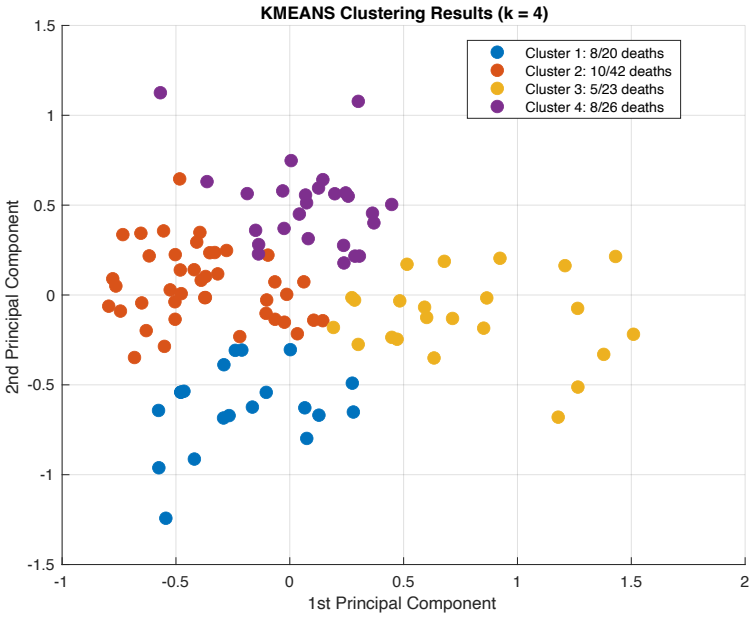


Figure 46 - The result of the grouping obtained with the K-means algorithm (k=4), with the number of deaths, displayed by the first two main components.

The division into four clusters obtained by the K-means method shows (Fig. 46) a different number and a different intrinsic mortality rate. The mortality rate of the groups is collected in Table 4.

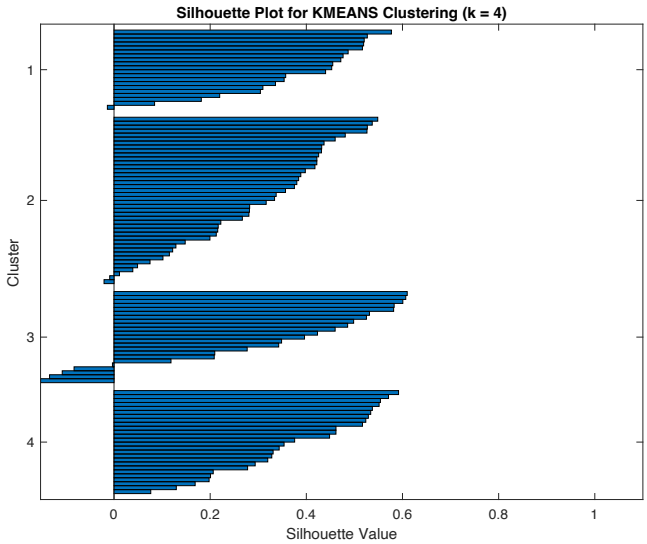


Figure 47 - Silhouette graph for the K-means algorithm (k=4)

The division, graphically shown by the silhouette graph (Fig. 47), shows four distinct groups, with 1 subject probably mistakenly assigned in cluster 2, 2 subjects in cluster 2 and 5 in cluster 3. The average silhouette value is equal to $\overline{S}(i) = 0.3339$.

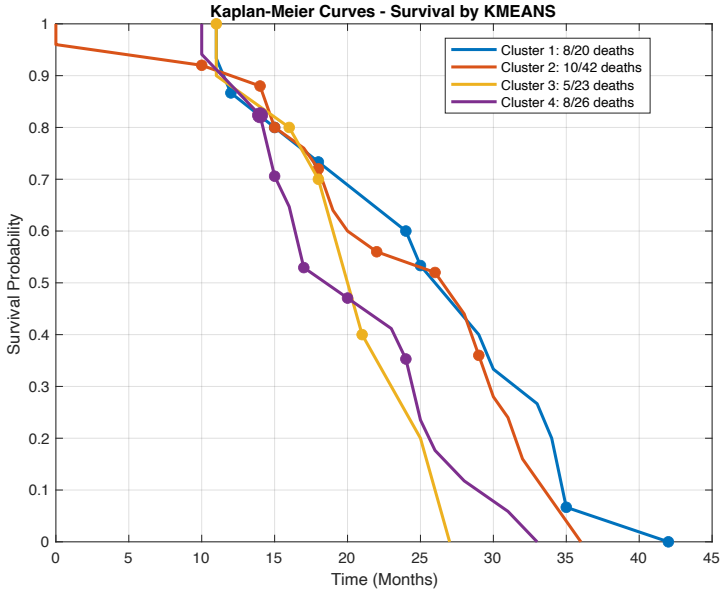


Figure 48 - Adverse event survival curve for the K-means algorithm (k=3)

Again, the event survival curves (Fig. 48) are not well distinguished.
In the initial part, before 15 months, all the curves are superimposed, showing various trend reversals between the various groups.
This could be a similar progression of the disease for all identified clusters.
In the following months, the curves are more separated, although there are various changes in trend, especially between group 1 and group 3 and between group 3 and group 4.
In addition, the curves identified do not reflect the prognosis in terms of death.
By analyzing the 40% probability of survival to the adverse event, we can see how the yellow curve (cluster 3) has a lower survival to the event than the blue curve (cluster 1).
However, the mortality rate in cluster 1 is the highest among the various groups, and that of cluster 1 is the lowest.
This result would indicate an inverse correlation between the probability of worsening or recurrence and the death of the subjects.

- Hierarchical clustering:

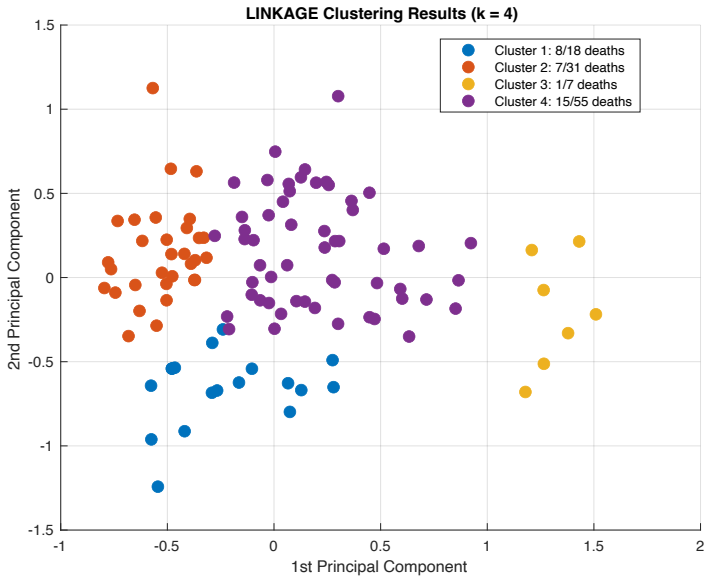


Figure 49 - Result of the grouping obtained with the linkage algorithm (k=4), with number of deaths, displayed by means of the first two main components.

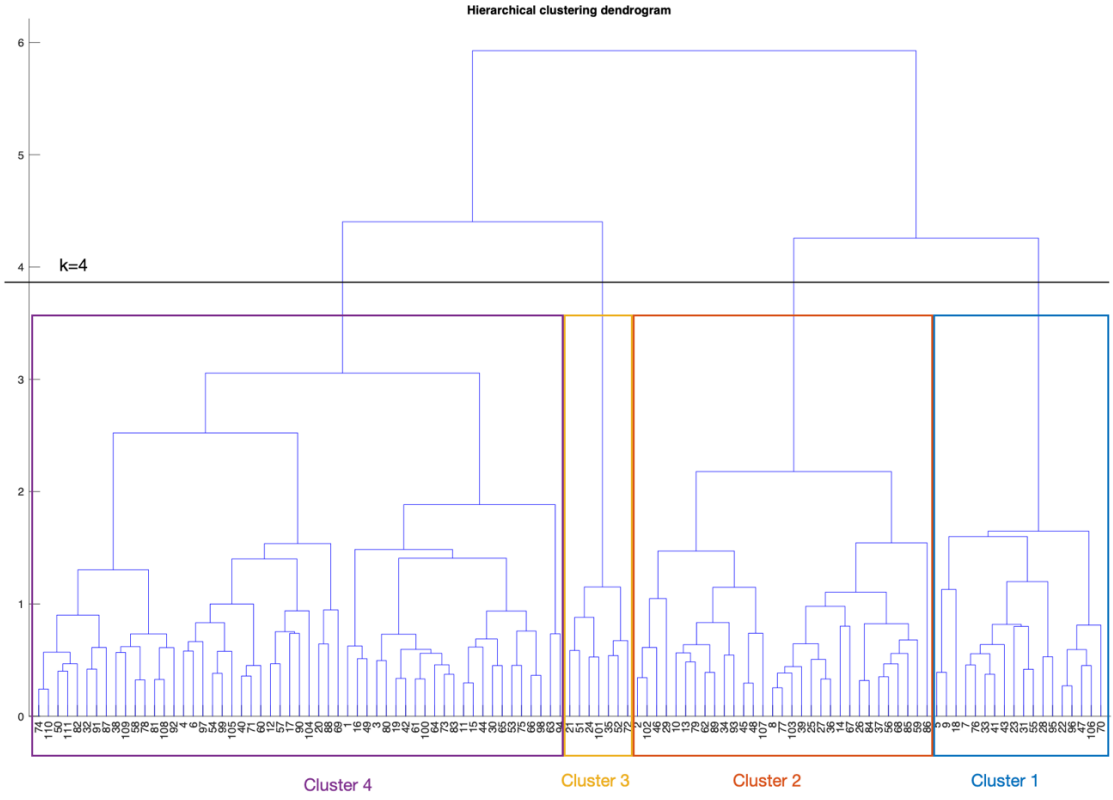


Figure 50 - Dendrogram representative of hierarchical clustering obtained from normalized features (k=4)

In Figure 49 it is possible to visualize the groups of patients created by the linkage algorithm, for k=4. Clusters 1 and 2 are about equally distant from clusters 3 and 4, as shown by the height of the segment that unites them in a hierarchical scale (Fig. 50). Cluster 1, in

blue, is also associated with the highest mortality rate, while cluster 3 is the group associated with the lowest mortality rate.

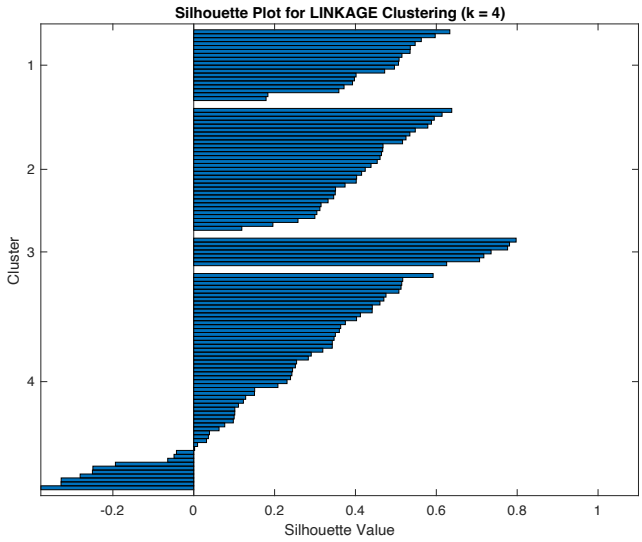


Figure 51 - Silhouette graph for the K-means algorithm (k=4)

The silhouette graph (Fig. 51) shows a good separation of the clusters for the first three groups, while in cluster 4 there are 10 subjects with possible belonging to the wrong cluster (negative silhouette). The average value of the silhouette is equal to $\overline{S}(i) = 0.3277$.

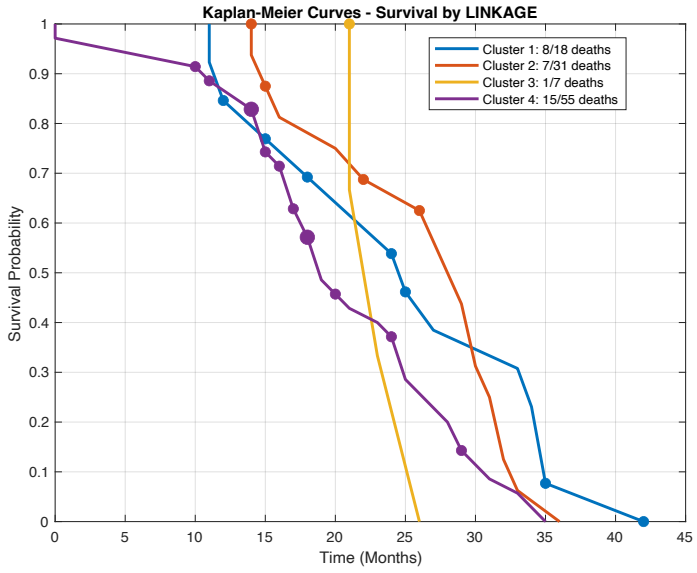


Figure 52 - Result of the grouping obtained with the linkage algorithm (k=4), with number of deaths, displayed by means of the first two main components

The adverse event survival curves obtained by the hierarchical clustering algorithm (Fig. 52) are more separated than the clusters obtained by the k-means algorithm, with the same number of clusters obtained.

Cluster 3, in yellow, seems to have a higher survival to the adverse event than the other curves, but an almost vertical trend in the course of the disease. Survival to cluster event 3, at least in the early stages, reflects the lower mortality rate associated with the group.
Cluster 4 and Cluster 1 show overlapping trends in the initial phase, showing distinct behaviors only after 15 months.
The trend of cluster 2, in orange, maintains its trend, reflecting its low intrinsic mortality, up to about 30 months, where it reverses its trend with cluster 1.

Table 4 summarizes the parameters that characterize the groups identified by the two algorithms tested with k=4.

Table 4 - Comparison of cluster parameters (k=4)

Method	*Clusters*	*Percentage of deaths*
Clustering K-means	Cluster 1	40.00%
	Cluster 2	23.81%
	Cluster 3	21.74%
	Cluster 4	30.77%
Hierarchical clustering	Cluster 1	44.44%
	Cluster 2	22.58%
	Cluster 3	14.28%
	Cluster 4	27.27%

From the use of machine learning methods for the stratification of a group of patients with DLBCL, the use of morphological features was found to be, in the case of hierarchical clustering with two groups, the best algorithm to discriminate subjects in terms of survival to the event.
The results obtained demonstrate, in the remaining cases analyzed, that the characteristics used to obtain the clusters are not sufficiently discriminating to reflect the survival of the adverse event, in terms of recurrences or aggravations of the disease.
In the case of subjects with DLBCL, in fact, the morphological features alone are generally highly heterogeneous and not very informative for the division into groups based on prognosis.

The result obtained demonstrates how the optimized extraction of characteristics, in terms of texture and shape variables, under the pathologist's indication, can still be effective, despite the intrinsic difficulty of the problem.

3.3 Critical issues and improvements

From the analysis of the results obtained, it is possible to deduce that the approach used for the stratification of patients is adequate to reflect the survival of the subjects. The result of the clustering carried out with the linkage method, with k=2, can represent, in fact, a starting point for the analysis of the populations affected by DLBCL. From the methods analyzed as a whole, however, it is evident the difficulty in the stratification of subjects affected by DLBCL, using morphological features alone.

It is possible that the data used within the feature matrix is connected by complex networks that may not have been captured by the tested algorithms (K-means and hierarchical clustering). However, it was not possible to use and compare other clustering algorithms on the available data because the data has the same average density and the variables are also linearly dependent.

ROIs obtained from histological data could, in a future study, be identified by a guided algorithm, which considers areas of clinical interest and automatically discards areas of non-interest. The algorithm to generate the ROI currently used draws 50 tiles randomly, pulling them from the entire WSI.

In addition, ROI staining could be standardized by a stain normalization algorithm to generate a dataset free of intra-patient and inter-patient color variability.

The data available to conduct the study are currently scarce and inadequate to be used within a clinical trial. It would be appropriate, in fact, to collect additional clinical data from patients, to allow a statistical study on the characteristics to be carried out and identify any significant parameters for the stratification of the populations. Such data could also be introduced into the matrix of features provided as input to the clustering algorithms, to allow the method to extract hidden patterns, possibly related to the morphological variables of the subject.

To improve the analysis, it would also be appropriate to collect data relating to further reports, in order to complete the subject's medical history and extract characteristics relating to the prognosis of the analyzed population.

In the study of cancer subjects, in fact, it is essential to consider various clinical, biological and molecular parameters for a complete evaluation of the disease. These parameters not only help diagnose the disease, but also define its prognosis and personalize treatment.

In patients with DLBCL, it is essential to collect genetic data to identify biological markers related to prognosis.

The expansion of the dataset and the characteristics associated with the subjects could allow the use of deep learning approaches, machine learning techniques more suitable for identifying the deep correlation between data, thanks to the intrinsic extraction of the characteristics linked to the output.

The comprehensive evaluation of a patient with DLBCL therefore requires a multidisciplinary approach that integrates clinical, biological, molecular and imaging information. This method is the best for defining the prognosis and planning treatment. Using an AI-based multimodal algorithm could lead to significantly better results for patient stratification and analysis of their survival.

The tested method and the characteristics extracted from the population of subjects analyzed is however a first starting point for the stratification of subjects affected by DLBCL, showing how the optimization of the extracted parameters can, even in the face of a heterogeneous group of subjects, allow the clustering of patients.

Conclusion and future developments

The study demonstrated how the application of artificial intelligence techniques for the stratification of patients with DLBCL, based on morphological characteristics, represents a potential tool to improve the understanding of the disease and clinical management. In particular, the use of hierarchical clustering allowed us to obtain two subgroups of populations, based on the extracted morphological characteristics that reflect the percentage of associated deaths and the prognosis of the subjects themselves.

However, the results obtained highlighted some limitations in the approach used, in the ability to accurately reflect patient survival. This is likely due to a complex and inherent data network that is not fully captured by the algorithms tested, as well as the scarcity of available data.

The difficulties encountered underline the need to improve the quality and quantity of the data collected and the need to use genetic and molecular information, to allow a more precise and useful stratification for the personalization of the treatment of the subjects, with a view to personalized medicine. A multimodal approach combining clinical, genetic, biological, and imaging data, supported by artificial intelligence, could make it possible to extract hidden patterns and significantly improve prognosis and treatment planning for patients with DLBCL.

In anticipation of an expansion of the available data, a multimodal algorithm has been developed, based on deep learning, capable of taking clinical, histological and genetic data as input and returning the survival curve of the patients analyzed as output. The algorithm was developed by adapting the '*Pathomic Fusion'* method developed by Chen et al. [24]. The algorithm uses various unimodal networks, VGG for histological image analysis, GCN for graph analysis, and SNN for molecular data analysis. The unimodal outputs are combined with each other to generate a complex network with optimized performance. In the process, Kronecker's product of unimodal features is used, combined with a gating-based attention mechanism to control the expressiveness of each representation. This helps you understand how the importance of features changes based on the input you provide. Finally, the output of the learning model can be used to estimate the probability of survival of the analyzed subjects.

Continuing research in the field of DLBCL is critical to better understand the mechanisms underlying the disease, identify novel prognostic biomarkers, and optimize

personalized therapies, improving the overall prognosis of the subject. The integration of advanced artificial intelligence techniques within clinical practice opens a promising avenue for clinician decision support in the management of cancer patients, contributing to the rapid processing of large data sets, facilitating patient management and aiming for precision medicine. Looking ahead, the use of AI algorithms in medical practice could revolutionize the ability to interpret large amounts of complex data, enabling the identification of hidden patterns and correlations between biological, clinical, and imaging features that would otherwise escape traditional analysis. Approaches based on machine learning algorithms can offer valuable support to traditional medicine.

In conclusion, continued investment in clinical research, in combination with the use of advanced technologies, will be decisive in achieving significant advances in the prognosis of patients with DLBCL and other complex and heterogeneous diseases that, to date, remain difficult to effectively treat with traditional approaches.

Bibliography

[1] Li S, Young KH, Medeiros LJ. *Diffuse large B-cell lymphoma*. Pathology. 2018 Jan; 50(1):74-87. doi: 10.1016/j.pathol.2017.09.006. Epub 2017 Nov 20. PMID: 29167021.

[2] Martelli M, Ferreri AJ, Agostinelli C, Di Rocco A, Pfreundschuh M, Pileri SA. *Diffuse large B-cell lymphoma*. Crit Rev Oncol Hematol. 2013 Aug; 87(2):146-71. DOI: 10.1016/J.Critrevonc.2012.12.009. Epub 2013 Jan 30. PMID: 23375551.

[3] Tsimberidou AM, Fountzilas E, Nikanjam M, Kurzrock R. *Review of precision cancer medicine: Evolution of the treatment paradigm*. Cancer Treat Rev. 2020 Jun;86:102019. DOI: 10.1016/j.ctrv.2020.102019. Epub 2020 Mar 31. PMID: 32251926; PMCID: PMC7272286.

[4] Di Napoli, A., Remotti, D., Agostinelli, C. *et al. A practical algorithmic approach to mature aggressive B cell lymphoma diagnosis in the double/triple hit era: selecting cases, matching clinical benefit*. Virchows Arch 475, 513–518 (2019). doi.org/10.1007/s00428-019-02637-2

[5] G. Wright, B. Tan, A. Rosenwald, and L. M. Staudt, *A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma*, Proceedings of the National Academy of Sciences, vol. 100, no. 17, pp. 9991-9996, Aug. 2003. doi: 10.1073/pnas.1732008100.

[6] *Digital pathology pathways to improve patient care*. Unpublished internal document.

[7] Vrabac D, Smit A, Rojansky R, Natkunam Y, Advani RH, Ng AY, Fernandez-Pol S, Rajpurkar P. *DLBCL-Morph: Morphological features computed using deep learning for an annotated digital DLBCL image set*. Ski Date. 2021 May 20; 8(1):135. DOI: 10.1038/S41597-021-00915-W. PMID: 34017010; PMCID: PMC8137959.

[8] Schmitz R, et al. *Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma*. N Engl J Med. 2018 Apr 12; 378(15):1396-1407. doi: 10.1056/NEJMoa1801445. PMID: 29641966; PMCID: PMC6010183.

[9] Wright GW, et al. *A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications*. Cancer Cell. 2020 Apr 13; 37(4):551-568.e14. DOI: 10.1016/J.ccell.2020.03.015. PMID: 32289277; PMCID: PMC8459709.

[10] J. S. Abramson, *Hitting back at lymphoma: How do modern diagnostics identify high-risk diffuse large B-cell lymphoma subsets and alter treatment?.* Cancer. vol. 125, no. 24, pp. 4292-4301. Jul. 2019; doi: 10.1002/cncr.32145.

[11] Del Giacco L, Cattaneo C. *Introduction to genomic*s. Methods Mol Biol. 2012;823:79-88. doi: 10.1007/978-1-60327-216-2_6. PMID: 22081340.

[12] Jiang Z, Zhou X, Li R, Michal JJ, Zhang S, Dodson MV, Zhang Z, Harland RM. *Whole transcriptome analysis with sequencing: methods, challenges and potential solutions.* Cell Mol Life Sci. 2015 Sep; 72(18):3425-39. doi: 10.1007/s00018-015-1934-y. Epub 2015 May 28. PMID: 26018601; PMCID: PMC6233721.

[13] Wu X, Fang Q**.** *Stacked autoencoder based multi-omics data integration for cancer survival prediction*. *ArXiv*. 2022; ABS/2207.04878.

[14] Manos J. *The human microbiome in disease and pathology*. APMIS. 2022 Dec; 130(12):690-705. doi: 10.1111/apm.13225. Epub 2022 May 6. PMID: 35393656; PMCID: PMC9790345.

[15] Vlachavas EI, Bohn J, Ückert F, Nürnberg S. *A Detailed Catalogue of Multi-Omics Methodologies for Identification of Putative Biomarkers and Causal Molecular Networks in Translational Cancer Research*. Int J Mol Sci. 2021 Mar 10; 22(6):2822. doi: 10.3390/ijms22062822. PMID: 33802234; PMCID: PMC8000236.

[16] Cai Z, Poulos RC, Liu J, Zhong Q. *Machine learning for multi-omics data integration in cancer*. iScience. 2022 Jan 22; 25(2):103798. doi: 10.1016/j.isci.2022.103798. PMID: 35169688; PMCID: PMC8829812.

[17] Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. *Integration strategies of multi-omics data for machine learning analysis*. Comput Struct Biotechnol J. 2021 Jun 22;19:3735-3746. doi: 10.1016/j.csbj.2021.06.030. PMID: 34285775; PMCID: PMC8258788.

[18] Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. *Introduction to Machine Learning, Neural Networks, and Deep Learning*. Transl Vis Ski Technol. 2020 Feb 27; 9(2):14. doi: 10.1167/tvst.9.2.14. PMID: 32704420; PMCID: PMC7347027.

[19] Steinbuss G, Kriegsmann M, Zgorzelski C, Brobeil A, Goeppert B, Dietrich S, Mechtersheimer G, Kriegsmann K. *Deep Learning for the Classification of Non-Hodgkin Lymphoma on Histopathological Images*. Cancers (Basel). 2021 May 17; 13(10):2419. doi: 10.3390/cancers13102419. PMID: 34067726; PMCID: PMC8156071.

[20] Neijzen D, Lunter G. *Unsupervised learning for medical data: A review of probabilistic factorization methods*. Stat Med. 2023 Dec 30; 42(30):5541-5554. doi: 10.1002/sim.9924. Epub 2023 Oct 18. PMID: 37850249.
[21] Malathi L, Amsaveni R, Anitha N, Balachander N. *Reticuloendothelial malignancy of head and neck: A comprehensive review*. J Pharm Bioallied Sci. 2015 Apr; 7(Suppl 1):S145-57. doi: 10.4103/0975-7406.155867. PMID: 26015695; PMCID: PMC4439655.
[22] Wei L, Gan Q, Ji T. *Cervical cancer histology image identification method based on texture and lesion area features*. Comput Assist Surg (Abingdon). 2017 Dec; 22(sup1):186-199. doi: 10.1080/24699322.2017.1389397. Epub 2017 Oct 16. PMID: 29037083
[23] The MathWorks, Inc., *MATLAB* version *R2023b Documentation*, Natick, Massachusetts: The MathWorks, Inc., 2023
[24] Chen RJ, Lu MY, Wang J, Williamson DFK, Rodig SJ, Lindeman NI, Mahmood F. *Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis*. IEEE Trans Med Imaging. 2022 Apr; 41(4):757-770. doi: 10.1109/TMI.2020.3021387. Epub 2022 Apr 1. PMID: 32881682; PMCID: PMC10339462.