POLITECNICO DI TORINO

DAUIN - DEPARTMENT OF CONTROL AND COMPUTER ENGINEERING

Master Degree course in Computer Engineering

Curriculum
Data Science

A Deep Learning approach to integrate histological images and DNA methylation values

Supervisors
Prof. Elisa Ficarra
Eng. Marta Lovino
Eng. Francesco Ponzio

Candidate
Margheret Casaletto
ID: 250725

Academic Year 2020-2021

E tutto insieme, tutte le voci, tutte le mete, tutti i desideri, tutti i dolori, tutta la gioia, tutto il bene e il male, tutto insieme era il mondo. Tutto insieme era il fiume del divenire, era la musica della vita.

Hermann Hesse

Acknowledgements

The Bioinformatics course allowed me to discover a research field inspiring, enabling me to improve the expertise gained during my master's degree in the biological field. I want to express my appreciation to the people who made it possible to carry out this thesis work. I am grateful to my supervisors, Professor Elisa Ficarra and her collaborators, Marta Lovino and Francesco Ponzio. Besides the knowledge, the suggestions and the insights they shared with me, I particularly appreciated their constant helpfulness and human aspect.

Summary

Thanks to the development and continuous enhancement of new medical and biological technologies, Artificial Intelligence (AI) holds grand hope to revolutionize cancer malignancy detection in the biomedical field. One of the recently emerging challenges consists of developing Machine and Deep Learning (ML, DL) based frameworks involving the integration of heterogeneous data gathered from omics such as Genomics, Epigenomics, Transcriptomics, or Proteomics and biomedical images. Highlighting the inter-relationships between data of different nature may help to better understand the progression of complex diseases in order to obtain a more precise diagnosis and prognosis, providing the patient the best possible attainable therapies.

This work aims to investigate the integration between two data types: a specific category of biomedical images, the histological ones, and DNA methylation. The latter can reveal transcriptional regulation mechanisms and, consequently, it is suitable to study pathological conditions, particularly cancer. In this thesis, I consider colon cancer data derived from patients in The Cancer Genome Atlas (TCGA), one of the largest available repositories for this type of information. Concerning images, I also exploit an additional set of Regions Of Interest (ROIs) derived from an external dataset of colon cancer histological images, pre-cleaned and labeled in a previous study according to the presence of healthy or tumor tissue.

To achieve the aim, I train an image classification model to predict the malignancy in the histological images. Afterward, I analyze how methylation data affects the prediction performances by exploiting the correlation between the features extracted from the two data types.

The input data consists of the methylation samples, divided between healthy and tumor class, and the images, which are also globally labeled as tumor or healthy. In the preliminary phase of the work, I perform a division into train

set and test set for both data types, taking care to integrate both the image and methylation data for the same patient. Next, I develop two pipelines in parallel that perform the same tasks for the two data types, exploiting an ML/DL approach based on the distinct nature of the data.

Regarding methylation, after a preprocessing step, I train multiple genomic classifiers and analyze the prediction scores on the test set. All the trained genomic classifiers achieve an accuracy higher than 94%. At this point, I evaluate two dimensionality reduction techniques, Principal Component Analysis (PCA) and Autoencoders (AE), to extract different feature sets from the methylation train set. Therefore, I train a Support Vector Machine (SVM) classifier for each extracted feature set and choose the feature set that achieves the best scores on the test set.

On images, the preprocessing step involves cutting the whole images and ROIs into smaller crops. For images sourced from the TCGA repository, I also handle background removal. I exploit a well-known Convolution Neural Network (CNN) architecture, the VGG16, to develop the image feature extractor model. After a hyper-parameters tuning procedure, I perform a complete VGG16 fine-tuning on the ROIs. I evaluate a second model by performing a further complete fine-tuning on part of the TCGA train set. The CNN is the first part of a features extraction pipeline that eventually performs PCA to obtain the same number of features extracted from the methylome data. I classify the test set images with both models and obtain two baseline results. I extract the train set features with both models, evaluating different feature sets, and train a Multi-Layer Perceptron (MLP) for each feature set. I choose the MLP that classifies the test set more likely to the respective baseline, hence making the extracted features representative. The selected MLP becomes the actual images baseline classification model.

I perform the integration between the features extracted from the image and methylation data for both the train and the test sets, exploiting two different statistical methods: Mutual Information (MI) and Pearson correlation. In detail, for each of the crops belonging to each whole slide, I have a vector of extracted features; even if one image is divided into multiple crops, it belongs to a single patient. Instead, in the methylation dataset, each patient is associated with a single sample, and therefore a single vector of extracted features is available. For this reason, the correlation is performed between all the crops of a specific patient with his methylation data. The crops belonging

to the tumor images are correlated with the corresponding tumor methylation sample, similarly for the healthy data. The MI method returns only non-negative values; results equal to 0 indicate that the feature vectors are independent of each other. Instead, Pearson's correlation between the two vectors yields outcomes in a range [-1, 1]. These values are used to discard all those image crops with a correlation value below a certain threshold: in the case of Mutual Information, I choose a threshold value equal to 0; as for Pearson, I discard crops that have a correlation value below the first quartile (25%) of the maximum correlation value.

In the final part of the work, I compare the prediction results from the baseline image classification model with those obtained from the same predictive model, but without accounting for the under-threshold crops described above. Therefore, I obtain three different sets of prediction scores on the test set (Baseline case, MI case, and Pearson case). Assuming that a slide is globally labeled as tumor if at least 10% of the crops is labeled as tumor, I conclude that the MI-based approach is the best.

The main challenges of this analysis mainly derive from the image data coming from the TGCA repository. Although the database provides a global label for each image, there is often a non-negligible percentage of other tissues inside (e.g., stromal tissue), which adds noise and introduces an error in the training of the models. In detail, it would be necessary to have at least a third-class available to distinguish between healthy, tumor, and other tissue types to improve the reliability of the results. It could improve the performance of the feature extractor model and consequently of the correlation values.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Artificial Intelligence (AI) has contributed to the progress of diverse research fields in recent years. Machine Learning (ML) is an AI application that aims to develop algorithms able to automatically learn from data and provide experience-enhanced predictions without being explicitly programmed. In turn, Deep Learning (DL) is an ML subcategory that has found remarkable success in multiple areas (computer vision, natural language processing, bioinformatics, etc.) for its ability to identify features of data automatically at different levels of abstraction.

In the biomedical area, thanks to the development and continuous enhancement of new medical and biological technologies, AI holds grand hope to revolutionize cancer malignancy detection. One of the challenges recently emerging is the development of ML and DL based frameworks that involve the integration of heterogeneous data gathered from omics studies and biomedical images. On one side, the advent of Next Generation Sequencing (NGS) technologies has made it possible to perform massive DNA and RNA sequencing, producing billions of nucleotide sequences in a relatively short time. On the other side, biomedical imaging coming from various imaging technologies is already widely used to discover tissue and organ characteristics associated with particular pathological states. Highlighting the inter-relationships between data of different nature may help better understand the progression of complex diseases and obtain a more precise diagnosis and prognosis, providing the patient the best possible attainable therapies. As a relatively recent field of research, there are still few experiments in the literature, yet they are steadily increasing. In many of them, the researchers exploit the integration between some of the omics like Genomics, Epigenomics, Transcriptomics, or

Proteomics, and image data generated using different technologies such as Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), or histological images [1] [2]. The contribution of AI is essential to highlight possible connections between the different types of data that would be difficult to identify otherwise. Approaches employed often involve the development of ML and DL models able, for example, to identify imaging-derived features that provide information about underlying tumor biology, to improve the accuracy of disease classification or patient survival time prediction.

The challenges facing this interdisciplinary field are not trivial. Firstly, the necessity to have high quality and sufficiently large data; this is a hyper-present issue in the data science world: to create reliable models, the collected data (the raw material) should have high standards. Furthermore, since we are dealing with sensitive data, there are confidentiality requirements to consider, which add a further obstacle to data usage. Secondly, the need to have correct annotations, including proper data labeling, requires knowledge of the application domain. Finally, designing and testing diverse network models is needed to reveal meaningful relationships among the different types of data. It is fundamental to keep all these aspects in mind to build robust ML or DL models and thus trust AI in clinical applications.

1.1 Thesis objective

The main objective of this thesis is to investigate the integration between one specific category of images, the histological images, and DNA methylation (Figure 1.1). The latter can reveal transcriptional regulation mechanisms and, consequently, it is suitable to study pathological conditions, particularly cancer.
I use colon cancer patient data, freely accessible online on the Genomic Data Commons (GDC) Data Portal [3], a research program of the National Cancer Institute (NCI) that makes available data and information on cancer patients from some of the complete cancer genomic repositories. Concerning images, I also have an additional set of Regions Of Interest (ROIs) deriving from a set of colon cancer histological images available on the University of Leeds Virtual Pathology Project Website [4], cleaned and labeled in a previous study, based on the presence of healthy or tumor tissue [5].
More specifically, I implement two parallel pipelines for both types of data coming from GDC Data Portal able to (i) preprocess the data, (ii) extract

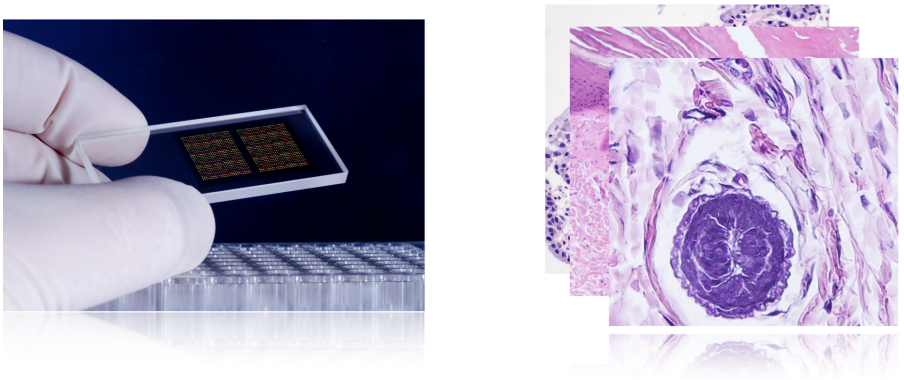

Figure 1.1. On the left, microarry technology used to obtain genomic values such as DNA methylation, On the right, an example histological image patches.

the features, (iii) validate the extracted features, adopting an ML/DL based approach. Finally, I build an image classification model to predict the whole image as healthy or malignant. Data integration is accomplished by correlating features extracted from images and DNA methylation. The final goal is to analyze how the correlation data impacts the image classification model, considering various evaluation metrics.

1.2 Organization

The work is distributed into the following chapters:

- Chapter 2: this chapter introduces the biological background, providing an overview of colon cancer, histological imaging, and DNA methylation. Successively, it presents an overview of some of the strategies found in the literature regarding the integration of omics and imaging using an ML/DL based approach.

- Chapter 3: This chapter represents the heart of the work. It describes the structure of the data and the preliminary operations to prepare them. Next, it focuses on the experiments performed for feature extraction and validation and creating the image classification model. Finally, it describes the integration methods examined and how these methods impact the aforementioned predictive image model.

- Chapter 4: this chapter presents the results derived from the different experiments conducted.
- Chapter 5: this chapter focuses on describing and interpreting the achieved results.
- Chapter 6: this chapter focuses on deriving a conclusion for the whole thesis.
- Appendix: this section contains supplementary material beneficial for understanding the effects of the method used, showing for some example slides the performed analysis.

Chapter 2

Background

This chapter introduces the biological background, providing an outline of colorectal cancer, DNA methylation, and histological imaging. After that, I present an overview of some of the approaches found in the literature regarding the integration of omics and imaging based on ML or DL.

2.1 Biological context

2.1.1 Colorectal cancer

The estimates of cancer incidence and mortality produced by the International Agency for Research on Cancer across over 20 global regions ranked colon cancer as the fourth most diagnosed cancer type in 2018 [6].

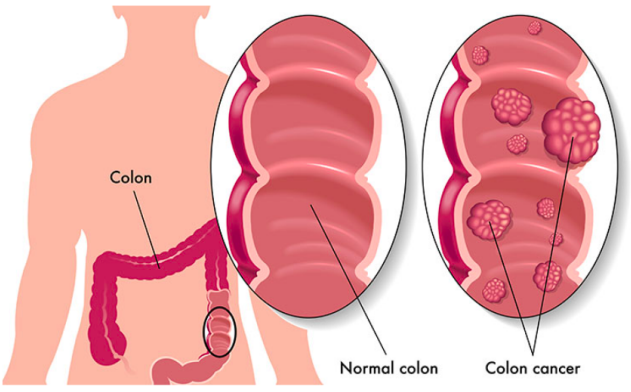

Figure 2.1. Colon position in the lower digestive system, shown in a healthy condition and affected by cancer [7].

Colorectal cancer (CRC) develops from the colon or rectum (part of the large intestine). In most cases, the tumor originates from small polyps that sit inside the intestine, evolving into cancer over the years (Figure 2.1). Possible risk factors may be having a family history of experiencing the disease (in two or more first-degree relatives) rather than having a personal medical history that may trigger the development of the malignancy (i.e., Crohn's disease). Other potential causes include alcohol abuse, smoking, and obesity. Possible diagnostic methodologies include sigmoidoscopy or colonoscopy. Once cancer is diagnosed, the prognosis and subsequent treatments depend on the stage of the disease. The first area in which the malignancy occurs is called a primary tumor (stage 0). When it begins to spread to other parts of the body (using the blood or lymph system), it is known as metastasis. The most common treatment to fight colon cancer involves surgery, removing the piece of intestine affected by the tumor [8].

2.1.2 DNA methylation

Epigenetics, as a branch of genetics, is concerned with studying gene activity. Nonetheless, it focuses on any changes resulting from gene expression not caused by modification of the DNA sequence [11]. An enlightening example concerns monozygotic twins, who share the same genotype; despite this, it may occur that they are not identical and, specifically, do not share the same epigenome [12]. Epigenetic processes are part of the organism's natural mechanisms and can be various. For instance, they may yield modification of chromatin (composed of the DNA and the proteins originating chromosomes) in a way that alters its structure and thus affects gene expression [13]. It is clear that studying epigenomics, with their alterations, leads to comprehending even some pathologies better, including cancer. The most known and studied epigenetic process is DNA methylation. As early as 1983, a study analyzed a reduction of DNA methylation for a specific gene in cells invaded by colon cancer [14]. Over the years, numerous studies have confirmed the relevance in observing DNA methylation as relating to cancer origin.

DNA methylation entails attaching a methyl group ($-CH_3$) to carbon 5 of the cytosine ring preceding guanine (CpG dinucleotides). The process is accomplished by a set of enzymes named DNA methyltransferases (DNMTs) that allow transferring the methyl group from S-adenosyl-methionine to cytosine. With the term CpG island (CGIs), it is meant all those areas in the DNA

that contain a significant presence of such CpG dinucleotides (cytosine followed by guanine). In reality, the amount of CpG islands inside the genome is not very high. They are usually present in the promoter or first exon region. Under normal conditions, in areas with unmethylated CpG islands, gene transcription is carried out, whereas, in areas with methylated CpG islands, transcription is inactivated (Figure 2.2) [15].

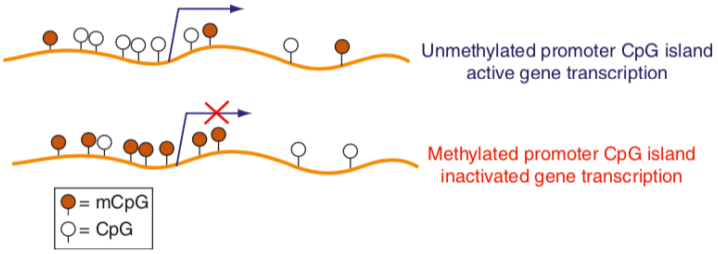

Figure 2.2. DNA methylation role in the activation or non-activation of gene transcription [15].

Therefore, undoubtedly, DNA methylation has a crucial function in the transcriptional regulation of genes, with a remarkable impact on genome stability. Abnormal conditions recur in two categories: hypermethylation and hypomethylation, indicating the excessive or poor presence of DNA methylation; respectively, they consequently lead to alterations of gene transcription. Several studies reveal how both conditions can be interrelated to the occurrence of cancer, demonstrating that cancer cells exhibit a different amount of methylation than normal ones. Generally, a state of hypomethylation is observed in cancer. On the other hand, hypermethylation of CGIs may be responsible for silencing specific genes in regulatory regions, which are not methylated in normal tissues. In addition, each cancer type may be associated with a specific methylome, meaning that methylation alterations are cancer-specific. Microarray technology makes it possible to examine DNA methylation values; thus, analyzing the amount of methylation in a cancer tissue can lead to early diagnosis of the disease and, consequently, essential in identifying a treatment response.

2.1.3 Histological images

Histology is concerned with studying the structure of cells and tissues, whose functions and characteristics are analyzed through microscopic investigation.

Histopathology focuses on studying tissues affected by the disease by examining a biopsy or surgical tissue specimen.

The digitized histological image is obtained after an initial preparation step of placing the tissue piece on a glass slide with chemicals or by freezing the slices. Pigmentation techniques are then employed to detect the different cellular components present within the tissue. Hematoxylin-Eosin (H&E) is undoubtedly the most widely used staining method. Other approaches involve the use of antibodies (immunohistochemistry, IHC) or immuno-fluorescence labeling. At this point, the slide can be scanned and thereby becomes computer analyzable [31]. The resulting whole slide images (WSIs) may have strong color differences caused by different scanning rather than staining procedures. Moreover, they can be explored at different levels of resolution, depending on different requirements; the higher the resolution level, the more in-depth inspection of histological features is.

Histopathology has long allowed pathologists to comprehend a disease state and decipher its progress, as well as confirm a diagnosis.[32]. The human decision support offered by analyzing histological images with the latest AI methods has been a hot topic in recent years. In reality, digital images appeared in the '80s, but their usability was hampered by the slowness of scanning equipment and a still scarcity of technological resources that today we are familiar with. A WSI can be around a dimension of 2 or 3 GB, which is excessive for the memories of those years. A significant first diffusion occurred from the following decade, in the 1990s, through moderately priced access to digital cameras, memory, and network resources [33].

Histopathology aims to trail what is already being accomplished by computer-aided diagnosis (CAD) algorithms, assisting the radiologist in patient prognosis. Similarly, it is possible to leverage technology to assist a pathologist in analyzing tissues within WSIs and potentially make disease classification easier [34]. Machine Learning is revolutionizing image processing, allowing WSIs to be analyzed automatically and harness the full power of their high resolution. Even if experienced in a particular domain, it is complicated for a human being to extract information from an extensive dataset. The MI/DL based approaches come in handy in revealing associations within the biological tissues that would otherwise be difficult to identify [35]. Image preprocessing is necessary to identify from the WSI the so-called Regions Of Interest (ROIs), i.e., areas of tissue consisting of similar features that can

therefore be univocally labeled. Undoubtedly the power provided by Deep Learning, specifically Convolutional Neural Networks, allows lightening a task such as the one just described, which otherwise would be excessively time-consuming. Specifically, thanks to the CNNs architecture, it is possible to extract various feature sets at different hierarchy levels (from less to most specific). While supervised algorithms make it possible to handle the classification of a particular disease (leveraging the power of neural networks as well as using classical Machine Learning algorithms such as KNN, SVM, or RF), through the unsupervised algorithms, it is possible to extrapolate different feature sets or clusterize data according to certain similarities. It is apparent that AI capabilities hold great promise for improving disease diagnosis and prognosis.

Major issues encountered in histological image analysis include:

- Insufficient or reliable labeled images. Although a global label of the whole image is provided, the more useful information is typically at patch levels, which is often unavailable.
- Huge image size that demands high memory resources.
- Magnification levels that imply different levels of information.
- Color variation due to multiple staining methods.

2.2 Machine and Deep Learning for genomics and images integration

This section describes some of the literature approaches that stand in the context of this thesis work. As announced in Chapter 1, this is a relatively emerging and exciting field of research, with a massive range of possibilities to span. Therefore, it is interesting to understand how researchers are approaching it and the main strategies employed.

Dai Yang et al. [48] present an Autoencoder-based method, designed to analyze the link between single-cell RNA-seq and chromatin images, aiming to identify distinct subpopulations of human T-cells ready for activation (that is critical to understanding the immune response). The method leverages AE technology to integrate and translate the two data types, mapping each

data type to shared latent space. Autoencoders are trained independently and then combined to translate between the different domain pairs. Thereby, they furnish a methodology for predicting the genome-wide expression profile of a particular cell given its chromatin organization and conversely. It is valuable for inferring how features in one dataset translate into features in the other. It is found that classifiers trained to discriminate among subpopulations in the initial datasets also performed well when assessed on the translated datasets.

Sun et al. [49] integrate genomic data (gene expression, copy number alteration, gene methylation, protein expression) and histological images from breast cancer patients. They propose a method called GPMKL, based on multiple kernel learning (specifically simpleMKL) that performs feature fusion coming from the different datasets embedded into cancer classification. The CellProfiler tool is used to extract features from the images. They focus on improving the prediction accuracy of breast cancer survival time by leveraging the information (the features extracted) coming from all types of data.

Smedley et al. [50] explore the associations between gene expression profiles and tumor morphology in magnetic resonance (MR) images of glioblastoma (GBM) patients. In detail, they train a deep neural network with both types of data to predict tumor morphology features. They exploit an approach based on transfer learning, initializing the weights (only some layers) of such network with the ones of autoencoder trained only on the genomic part. Comparing the results of predicting tumor morphological features with those obtained by linear regression proves their model achieves lower error levels.

Zhu et al. [51] propose a framework for predicting lung cancer survival by exploiting gene expression data (RNA) and histological imaging. They extract the most relevant features from the two data separately and then integrate them to train the survival model. The results show that the molecular profile information and pathological image information are complementary and, more importantly, demonstrate the improved prediction performance of the proposed integration compared to using only genomic or imaging data.

Chen et al. [52] propose a general paradigm that not only predicts cancer survival time but can also improve patient stratification. To validate their method, they use glioma and clear cell carcinoma datasets, for which they have genomics (CNV, mutation status, RNA-seq expression) and histological images. They extract in parallel two different feature sets from the images: one based on CNNs and the other exploiting Graph Convolutional Networks (GCNs), while the genomic features are extracted using Self Normalizing

Networks (SNNs). They build an ad hoc mechanism to check each extracted feature set's expressiveness and then exploit the Kronecker product to model the pairwise feature interactions from the different networks. They highlight the high interpretability of their paradigm, which allows one to understand which features (from the different data sets) are factored into predicting the survival outcome.
Mobadersany et al. [53] develop a deep learning-based model, named Genomic Survival Convolutional Neural Network (GSCNN), for survival prediction in brain cancer. In the model architecture, convolutional layers are succeeded by a sequence of fully connected layers. The framework embodies genomic data as inputs to the fully connected layers. The last layer enables the modeling of the survival estimation event. Throughout their model, they show prediction accuracy that exceeds the current clinical paradigm for predicting overall survival of patients diagnosed with glioma.
Hao et al. [54] aim to integrate GBM histological images and gene expression using a DL based approach. Their purpose is to improve survival prediction, including discovering genetic and histopathological patterns that may result in different survival rates in patients. They leverage a deep learning approach, building an architecture that identifies survival-related features without having hand-labeled ROIs available. Furthermore, they provide a patch-level feature aggregation strategy to obtain global features.

Unquestionably, putting together imaging data and omics is not a trivial task. The approaches described focus primarily on the use of histological images, but if one were to consider all categories of bioimaging combined with the various omics, the variability is large. Therefore, the proposed solutions are commonly based on feature extraction from the different types of data and subsequent integration to summarize the reported integrative methods. Besides the standard classification (tumor/non-tumor) and survival time prediction, the paradigms are often designed to be biologically interpretable, thus allowing a better study of the heterogeneity of the tissues present in the images (with a more comprehensive targeting of cell types), rather than seeking to improve patient stratification, treatment response, and treatment resistance.

Chapter 3

Method

In this chapter, after an accurate description of the data structure, I focus on the preliminary data preparation operation. First, I analyze the methylation dataset, showing that the genomic classifiers trained on it are strong and reliable. Then, I describe the methods used to extract the features and to validate them.

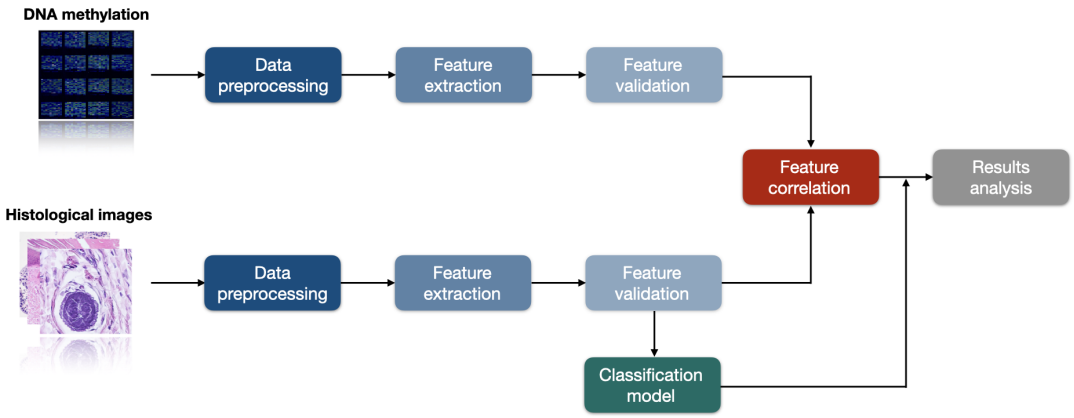

Figure 3.1. High level workflow. After a preliminary preprocessing phase, the most significant features are extracted and correlated with each other. The ultimate goal is to analyze how the integration between the two data types impacts the results of the image classification model.

As already anticipated in Chapter 1, the steps of preprocessing, feature extraction, and feature validation are performed, albeit with different approaches, for both data types (Figure 3.1). In this way, I describe how I realize the image classification model and where I exploit the correlation between the extracted features from the two data types to make changes to the prediction results of that model.

3.1 Data description

The data to be integrated come from The Cancer Genome Atlas (TCGA), a well-known archive featuring over 20,000 primary cancer samples and matched healthy samples covering 33 cancer types. The data are freely accessible online at the GDC Data Portal [3], thanks to the efforts of a research program of the National Cancer Institute. As for the images, I also have an additional set of ROIs derived from a set of colon cancer histology images available on the University of Leeds Virtual Pathology Project website [4], cleaned and labeled in a previous study according to the presence of healthy or tumor tissue [5].

An overview of available data follows:

- TCGA repository source:
  - Methylation samples. Each sample represents a specific patient for which the label (tumor or healthy) is provided.
  - Whole slide images. Each image represents a specific patient, for which a global label (tumor or healthy) is provided. Only for tumor slides can one get the percentage of tumor cells, healthy cells, or any other tissue, occurring in each slide.
- External source:
  - Region Of Interests. Each ROI carries an assigned label (tumor or healthy). As each of them stems from a Whole Slide, different ROIs may belong to the same patient.

In each case, I examine patients who have been diagnosed with colon cancer.

3.1.1 Methylation dataset

I collect data by selecting methylation beta values produced with Illumina Human methylation platform 450 and Illumina Human methylation platform 27 on the GDC Data Portal. I construct the dataset in such a way to have on the rows the patients and the columns the features (CpG islands).
The dataset has a size of (534, 25978). That is 534 patients by 25978 features. As can be noticed in Figure 3.2, it is strongly unbalanced between the healthy and tumor class: 460 tumor samples (86.14%) to 74 healthy (13.86%).

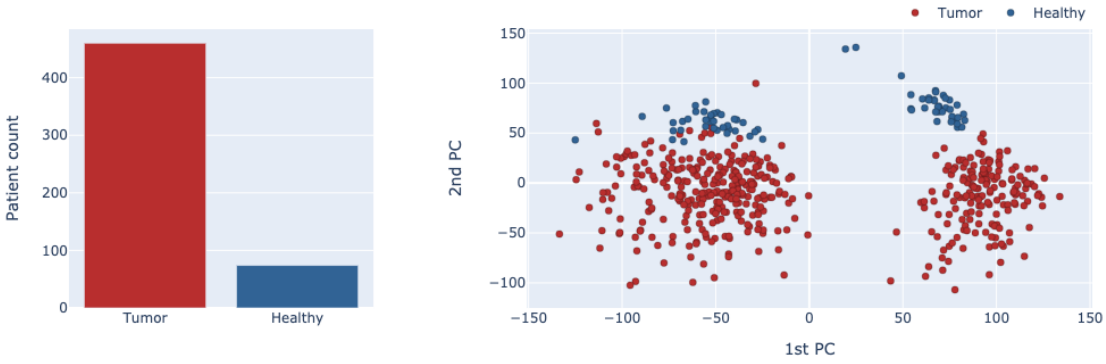

Figure 3.2. On the left, you can see the classes division in the methylation dataset. On the right, a visualization of the entire dataset using PCA is shown.

3.1.2 Images datasets

Regarding the data coming from the TCGA repository, I download slide images, considering tissue slides as an experimental strategy. As a sample type, I select "primary tumor" to collect images labeled as tumor; "solid tissue normal" to obtain images labeled as healthy.
Both the TCGA slides and ROIs are files saved in SVS format and therefore in multi-resolution. To open them, I use OpenSlide Python [59], a Python interface to the OpenSlide library that allows reading a small volume of image data at the resolution closest to a preferred zoom level.
The Table 3.1 shows, for the two datasets, the number of patients available and their division into tumor and healthy class. In TCGA images, one slide is associated with one patient, so I have 556 starting images. In the second

	Size	Class	
	# Patient	Tumor	Healthy
Whole slides	556	462	94
ROIs	18	9	9

Table 3.1. Number of patients for each set of images and relative class division.

case, several ROIs can belong to the same patient; therefore, I have 80 ROIs available, belonging to 18 distinct patients.

3.2 Data preprocessing

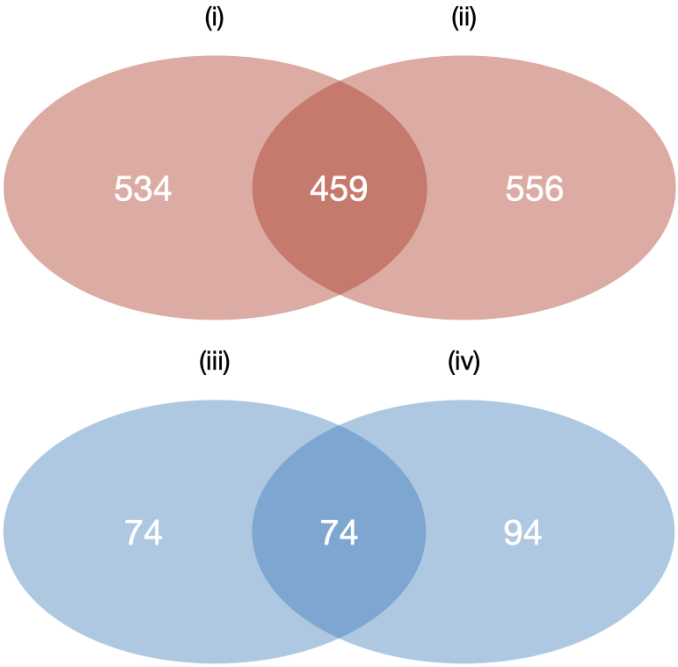

Figure 3.3. Venn diagrams show patients in common between the two TCGA datasets. (i) Number of tumor patients belonging to the methylation dataset.(ii) Number of tumor patients belonging to the image dataset. (iii) Number of healthy patients belonging to the methylation dataset. (iv) Number of healthy patients belonging to the image dataset.

As a preliminary action, I check TCGA patients for whom I have both the methylation and image data:

- Number of patients for whom methylation and images are available in the tumor class: 459.
- Number of patients for whom methylation and images are available in the healthy class: 74.

The Venn diagrams in the Figure 3.3 display the number of shared patients among the two types of data for the tumor class (red) and the healthy one (blue).

3.2.1 Train and test set preparation

Given the samples in common between the two TCGA datasets, I create a train set and a test set for methylation, and I do the same for the images. They address the following criteria:

- The methylation test set and the images test set contain the same patients.
- The methylation train set and the images train set also contain the same patients. The train set is generated from a subset of the remaining samples, making the two classes balanced.
- If both tumor and healthy data are available for the same patient, they are included in the same set.

External ROIs are used as an additional train set available.

The Table 3.2 shows the division of the samples between train and test set and the corresponding division in the two classes.

		Size	**Class**	
		# Patient	**Tumor**	**Healthy**
TCGA	Train set	86	43	43
	Test set	149	119	30
ROIs ext.	Train set	18	9	9

Table 3.2. Number of patients in train and test set and relative class division.

3.2.2 Methylation analysis

The methylation dataset contains real values in a range [0.0031, 0.995]. No 0 values are present, although there is a significant number of missing values. Since the number of features is high, I decide to remove all columns for which at least one value is missing (3839). The dataset shape after removing the missing values becomes (534, 22139). Next, the samples are divided into train and test set as described in the previous paragraph.

Genomic classifiers

It is necessary to prove that the methylation data works well in the classification task to achieve the stated goal. For this reason, I implement a pipeline that consists of:

- Scaling the data: standardize features by removing the mean and scaling to unit variance (Formula 3.1), in order to work with data at a common scale of values.

$$\bar{x}_i = \frac{x_i - \mu_i}{\sigma_i} \tag{3.1}$$

  Where μ_i is the mean value and σ_i the standard deviation of the i-th feature.

- Dimensionality reduction: reduce the number of features by applying the PCA technique so that the amount of variance explained by the selected components reaches 65%.

- Training of 4 different supervised classifiers I will indicate as genomic classifiers: K-Nearest Neighbors (KNN), Support Vector Machine, Random Forest (RF), and Multi-Layer Perceptron.

- Evaluating the prediction results produced by the 4 genomic classifiers on the test set using the following scores: accuracy, precision, recall, and f1.

To optimize the parameter search of each classifier during the training phase, I perform 5-fold cross-validation by dividing the train set in an 80%-20% proportion. The train/validation set splitting is made to ensure that patients with both tumor and healthy samples are available are part of the same set.

I report as follows a table of optimized parameters[1] for each classifier, as well as the best ones highlighted in gray.

K-Nearest Neighbors

metric	manhattan	euclidean	minkowski
n_neighbors	3	5	7

Table 3.3. Grid-search parameters for KNN

Support Vector Machine

kernel	rbf	poly	
C	0.1	1	10
gamma	scale	auto	

Table 3.4. Grid-search parameters for SVM

Random Forest

criterion	gini	entropy

Table 3.5. Grid-search parameters for RF

Multi Layer Perceptron

learning_rate_init	0.01	1e-03

Table 3.6. Grid-search parameters for MLP

Nevertheless, I cannot assume that all the predictions will be correct. I have to take into account that the models are subject to prediction errors. Therefore, a so-called confusion matrix (Figure 3.4) can be considered for each classification model: TP (True Positives) and TN (True Negatives) are respectively the quantity of tumor and healthy samples that are correctly

[1]For any clarification on the meaning of the optimized parameters consult the scikit-learn library [60], the tool used to implement the experiments.

predicted. Analogously, FP (False Positives) and FN (False Negatives) are respectively the quantity of tumor and healthy samples that are wrongly predicted. In the ideal case, FP and FN are zero.

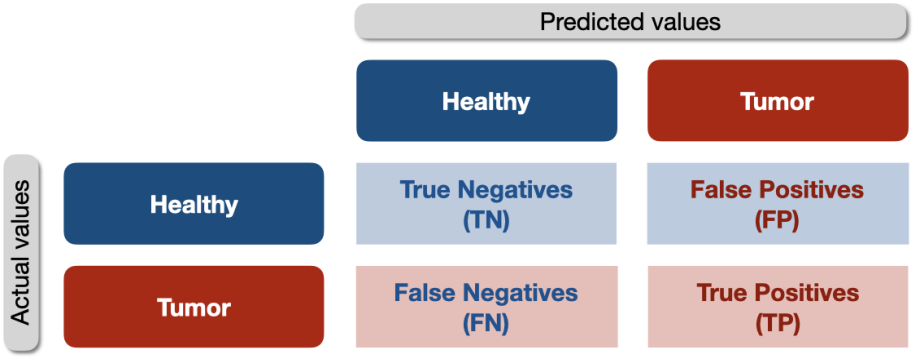

Figure 3.4. Confusion matrix.

From the confusion matrix I easily derive some useful predictive performance scores for each model: accuracy, precision, recall and f1-score. In detail:

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $f1 = 2 \cdot \frac{precision \cdot recall}{precision+recall}$

Accuracy is undoubtedly the most general measure for evaluating the classification performance of a model. However, other measures are frequently considered because they better show how the model performs on only one of the two classes. For example, to minimize FNs, it is appropriate to focus on recall. If, on the other hand, one wants to minimize FPs, precision is the most recommended. The f1 is defined as the harmonic mean between the two previous measures. Therefore, it allows obtaining a balance between precision and recall.

As noted in Table 3.7, all trained classifiers make predictions on the test set that achieve more than 94% accuracy. When looking at the other scores, the results are similar. It demonstrates that the methylation data allows for

the training of robust genomic classifiers, and it is, therefore, suitable for the idea of integration with the image data.

	KNN	SVM	RF	MLP
accuracy	94.6%	99.3%	99.3%	98.7%
precision	100%	100%	100%	100%
recall	93.3%	99.2%	99.2%	98.3%
f1	98.7%	99.7%	100%	100%

Table 3.7. Accuracy, precision, recall and f1 score for each genomic classifier.

3.2.3 Images: preliminary operations

Processing image data requires preliminary steps. The whole slides can reach huge dimensions since it is essential to inspect the tissues in detail. For this reason, the technology used to create those slides grants to obtain different levels of magnification and, therefore, read them at different levels. For example, it is possible to inspect the same image at a deeper level, taking advantage of a higher pixel resolution or at a lower level, depending on one's needs (Figure 3.5). With the previously mentioned OpenSlide library, images from both datasets (TCGA and external ROIs) are opened at level 0, the deepest level possible.

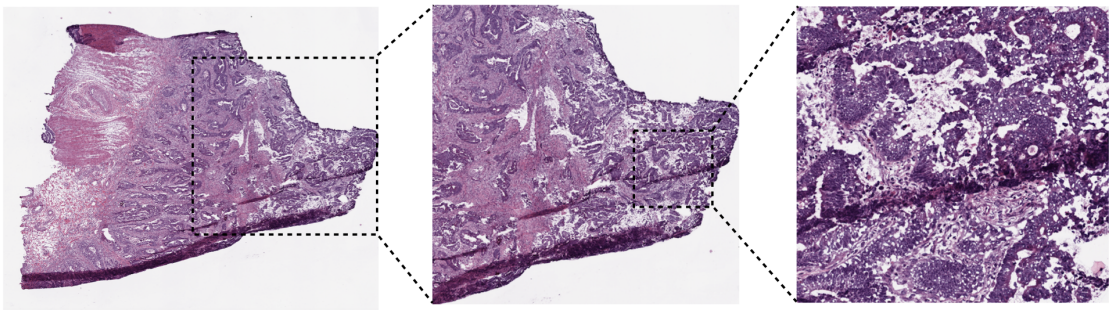

Figure 3.5. A whole slide example from the TCGA dataset showing the advantage of having multi-resolution images to inspect tissues at different zoom level.

Cropping

To build an image classifier model, whole slides need to be cut into smaller crops because of their high resolution. I develop a custom script in Python able to:

- Generate non-overlapping crops of 500x500 pixels size from each whole slide.
- Resize each crop to 64x64 pixels, using the thumbnail method provided by Python Imaging Library (PIL). With the before-mentioned method, obtaining a thumbnail version of a given image in a requested size is possible.
- Convert the RGBA image into an RGB one, to not consider the Alpha channel (it represents the degree of transparency).

In the end, each crop has a size of (64, 64, 3). The first two quantities represent the pixels array size, while the last one is the number of channels, which govern the color combination.

Background removal

As you can see from Figure 3.5, in TCGA whole slides, it is also necessary to deal with the background presence. The approach used foresees the discarding of all those crops whose average number of pixels is above a certain threshold, optimized based on different experiments. For example, given 255 color levels, for tumor slides, the chosen threshold is 232, while for healthy ones is 235. Even though many background crops are discarded with this approach, a background presence will likely remain in all crops cut on the boundary of the whole slide. Extracting crops accurately is a resource-intensive, mostly hand-crafted, and consequently time-consuming operation. So, you have to consider that the presence of background residue is a sure source of noise.
In the ROIs the background issue has already been managed in a previous study. So I only have to deal with cropping.

3.2.4 Final datasets

Table 3.8 summarizes the final size and relative class division for the methylation, images and ROIs datasets after the preprocessing step. Both patient-level and crop-level information are shown for images.

		Size	Class	
		# Patient / # Crops	Tumor	Healthy
Methylation	Train set	86	43	43
	Test set	149	119	30
Whole slides	Train set	86	43	43
		122018	77250	44768
	Test set	149	119	30
		339148	296726	42422
ROIs	Train set	18	9	9
		14623	7971	6652

Table 3.8. Final datasets size and relative class division.

3.3 Feature extraction

At this stage, for both types of data, I evaluate multiple models aiming to extract different feature sets, which will be successively validated.

3.3.1 Methylation

The goal is to find an acceptable trade-off between the quality and the number of features extracted from the methylation dataset. I consider two different approaches that both lead to reduce the dimensionality of the dataset: the PCA based and the Autoencoder[2] based one [62]. The former is nowadays intensively used for this kind of task while the latter is more challenging.

Principal Component Analysis approach

The flow is the following:

- Scaling the data: standardize features by removing the mean and scaling to unit variance in order to work with data at a common scale of values (Formula 3.1).

[2]the Autoencoder is a Neural Network composed of an Encoder block (mapping data in a latent space) and a Decoder block (mapping data in original space) built in a customized manner.

- Extraction of 3 different feature sets by performing a PCA, such that the number of selected components for each feature set achieves a percentage of explained variance of 65%, 80%, and 99%.

The Table 3.9 describes the number of features extracted for each feature extractor model exploiting PCA.

	% Explained variance	**# Feature extracted (PCA)**
Model 1	65	9
Model 2	80	23
Model 3	99	76

Table 3.9. Methylation: feature sets evaluated with PCA approach.

Autoencoder approach

The flow is the following:

- Scaling the data: normalize all features into a range [0,1], meaning that any given value may assume a value between the minimum value, 0, and the maximum value, 1. The mathematical formulation is as follows (Formula 3.2):

$$\bar{x}_i = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \tag{3.2}$$

  Where $min(x_i)$ and $max(x_i)$ are respectively the minimum and the maximum of the i-th feature.

- Feature selection by exploiting the chi-square statistical test, in order to restrict the number of initial features to be fed to the Autoencoder.

- Developing the Autoencoder architecture, assessing the following hyperparameters on the train set:

  – Loss function

  – Optimizer

  – Learning rate

  – Number of epochs

  – Batch size

- Autoencoder training. In this process, I employ a tool provided by Keras called EarlyStopping. It is essentially a callback that enables the monitoring of a selected metric, and it stops training when such metric stops improving. I keep a 10% percentage of the train set as a validation split. The callback monitors the validation loss to evaluate the number of useful epochs and prevent the overfitting effect.
- Extracting from AE the Encoder block that will be the feature extractor.

According to the pipeline outlined above, I create 4 different models, leading to 4 different feature sets. All models share these choices:

- The number of input nodes is equal to the number of features selected in the feature selection step.
- Use of Rectified Linear Units (ReLU) as non-linearities.
- Use of the batch normalization technique during training. It stabilizes the learning procedure and sharply reduces the number of training epochs, thus avoiding model overfitting.
- Use of a linear activation function in the output layer.

The Table 3.10 summarizes the hyper-parameters adopted for each Autoencoder.

	Loss function	**Optimizer**	**Learning rate**	**Epochs**	**Batch size**
AE 1	MSE	Adam	1e-4	60	16
AE 2	MSE	Adam	1e-4	40	16
AE 3	MSE	Adam	1e-4	40	16
AE 4	MSE	Adam	1e-5	60	16

Table 3.10. Hyper-parameters for each AE.

The Table 3.11 describes the number of features selected in the preliminary phase and the number of features extracted for each feature extractor model exploiting Autoencoders.

3.3.2 Images

The approach adopted in creating the feature extractor model for images is incremental. I first evaluate the model architecture, optimizing the search

	# Feature selected (FS)	# Feature extracted (Encoder)
Model 1	15000	20
Model 2	15000	150
Model 3	15000	512
Model 4	4000	20

Table 3.11. Methylation: feature sets evaluated with the Autoencoder approach.

for hyper-parameters on the train set of ROIs. Successively, after choosing the architecture and optimal parameters, I get two models: the first one is trained on the ROIs themselves; the second model is a further fine-tuning of the first model on a part of the train set derived from TCGA. From the models mentioned above, I derive the two possible convolutional-based feature extractor models in-pipelined to ultimately perform a PCA and reduce the number of extracted features more efficiently. I will subsequently validate those feature extractor models to determine which one should be used.

As a starting point, I evaluate two well-known CNN architectures in literature: VGG16 and ResNet50 [64] [65]. Typically, those CNN architectures provide two macroblocks for the sake of image classification:

- *Convolutional block*, mainly composed of the convolutional and pooling layers, which allows feature extraction.
- *Classifier block*, for classifying the images according to the discovered features.

It is a common practice to use pre-trained models on a huge reference dataset built to solve a similar problem to the one we would like to address; this leverages prior learning and avoids training the model from scratch. Specifically, I consider the two CNNs pre-trained on ImageNet, a huge dataset consisting of over 14 million images with associated labels, aiming to perform a fine-tuning on my work dataset. I exploit the implementation of the two CNNs made available by Keras [66], an open-source software library that provides a Python interface to many Convolutional Neural Network architectures described in the literature. Keras serves as an interface to the TensorFlow library.

After realizing the classification model eventually fine-tuned on the train set, it is possible to exploit it as a feature extractor by removing the classification block.

With this background, I realize a custom model in the following way:

- Select the CNN architecture model (VGG16 or ResNet50) pre-trained on ImageNet.
- Remove the classifier block included on the top.
- Adjust input data shape that the model expects according to the size of my working dataset.
- Add a preprocessing layer specific to the architecture selected.
- Choose the classifier to be placed on top of the convolutional macroblock. Rather than exploiting the standard approach of using a stack of fully connected layers, I add a global max-pooling layer.
- Add a dropout layer to avoid overfitting.
- Place the softmax layer at the end for classification.

To optimize the parameter search of each classifier, I decide to use the train set consisting of ROIs, which are finely cleaned and therefore less noisy than the images coming from TCGA. Thus, I perform 5-fold cross-validation by dividing the train set in a 90%-10% proportion. The train/validation set splitting is made to ensure that patients for whom both tumor and healthy samples are available are part of the same set and always guarantee that different patients belong to different sets.

I choose the following settings:

- Loss: categorical cross-entropy, since I use softmax activation function in the output layer.
- Batch size: 32
- Epochs: 20

I assess the following hyper-parameters on the train set:

- Network layer from which to start fine-tuning.

- Dropout rate
- Optimizer
- Learning rate. It is a good practice to use small learning rates in the fine-tuning process. Otherwise, the knowledge gained previously may be useless.

I report as follows a summary (Table 3.12) of optimized parameters[3]. I use the term "None" in the table to indicate that a complete fine-tuning has been performed. The last two columns represent the cross-validation scores (mean train score and mean test score) obtained with the best hyper-parameters, intentionally highlighted in grey.

Model	Layer FT	LR	Optimizer	Dropout rate	Mean train score	Mean test score
VGG16	None	1e-4	SGD	0.5	0.9528	0.8924
	11	1e-5	Adadelta	0.2		
	15	1e-6	Adam			
ResNet50	None	1e-4	SGD	0.5	0.9724	0.8738
	81	1e-5	Adadelta	0.2		
	143	1e-6	Adam			

Table 3.12. Hyper-parameters for each CNN architecture evaluated and scores achieved with best parameters.

Taking advantage of the results presented in the Table 3.12, I decide to use the model that leverages the VGG16 architecture, which proves to achieve a higher score on the test set than the ResNet50 based one. As anticipated, I train two image classification models that I will refer to as *Base model 1* and *Base model 2*. In Figure 3.6 you can visualize a block diagram of the architecture used to obtain the two models, in which I decide to insert also two additional layers for the data augmentation, always to prevent the overfitting. Specifically, I harness two Keras modules that are designed to perform a random flip of the input image horizontally and vertically and a counterclockwise rotation by a factor of 0.2.

Base model 1

The model is built by performing a complete fine-tuning of the architecture presented in the Figure 3.6 on the full ROIs train set, as established via the

[3]For any clarification on the meaning of the other optimized parameters consult the Keras library [66].

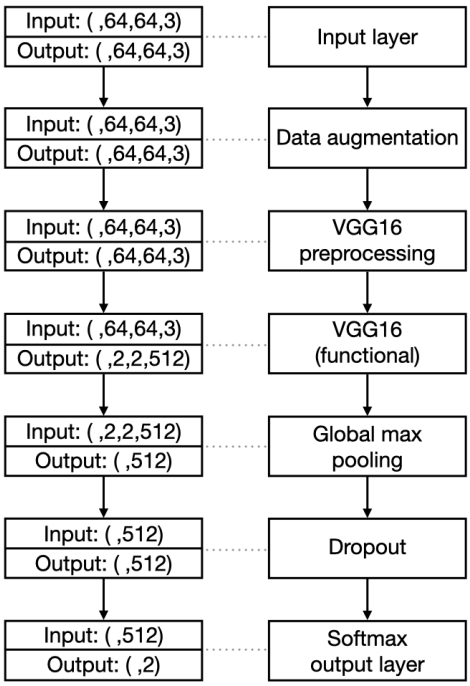

Figure 3.6. Image classification model block diagram and variation of the image size as it flows through the different layers. From this architecture I realize both Base model 1 and Base model 2.

cross-validation technique. Other parameters follow:

- Learning rate: 1e-6
- Optimizer: Adam
- Dropout rate: 0.2
- Loss: categorical cross-entropy
- Batch size: 32
- Epochs: 20

Base model 2

I perform a further complete fine-tuning of the Base model 1 on a part of TCGA train set (whole slides); for this reason, it is necessary to use a lower learning rate than those used to train Base model 1. As noted in Table 3.8, the train set is balanced in terms of patients, but after cropping, it turns

out that the tumor class is in a clear majority. I then perform a random undersampling, removing part of the crops belonging to the majority class to make a more balanced train set. The Table 3.13 shows the size and balance of the classes before and after undersampling.

		Size	Class	
		# Patients / # Crops	**Tumor**	**Healthy**
Whole slides	**Train set**	86	43	43
		122018	77250	44768
	Train set used	73	30	43
		91734	46966	44768

Table 3.13. Train set undersampling to build Base model 2.

The settings selected for training the model follow:

- Learning rate: 1e-7
- Optimizer: Adam
- Loss: categorical cross-entropy
- Batch size: 32
- Epochs: 90

Feature extraction pipeline

For either *Base model 1* and *Base model 2*, I carry out the following steps:

- Removing of the classifier block by cutting the model on the Global max-pooling layer. In this way, I get the convolutional-based feature extractor model that, starting from input data[4] of size (64, 64, 3) brings out a set of 512 features.
- Scaling the data: standardize features by removing the mean and scaling to unit variance in order to work with data at a common scale of values (Formula 3.1).

[4]Input data will be the individual crops belonging to each Whole slide.

- Extraction of 2 different feature sets by performing a PCA, such that the number of selected components for each feature set is equal to 9 and 20.

The Table 3.14 summarizes the feature sets evaluated using the procedure just outlined.

Pipeline	**# feat. extr. after CNN**	**# feat. extr. after PCA**
Base model 1 + PCA(n_comp=9)	512	9
Base model 1 + PCA(n_comp=20)	512	20
Base model 2 + PCA(n_comp=9)	512	9
Base model 2 + PCA(n_comp=20)	512	20

Table 3.14. Images: feature sets evaluated.

3.4 Feature validation

This stage aims to validate the feature sets extracted in the previous step, either for methylation or for images from TCGA, to be later integrated. First, I train some models using each extrapolated feature set to achieve the aim and classify the test set (both images and methylation). Then, depending on the results obtained, I decide which extracted feature set should be used for any two data types.

3.4.1 Methylation

It is needed to evaluate the extracted feature sets shown in the Table 3.9 and Table 3.11. I opt to train a Support Vector Machine per feature set since it is recognized as one of the most robust prediction methods. Like any classifier, it is essential in the training phase to optimize the hyper-parameters. Table 3.15 shows the parameter grid I select to optimize for SVM on each feature set.

For each feature set, the flow is:

- Optimization of the hyper-parameters shown in the Table 3.15 through a 5-fold cross-validation technique, maximizing the f1-score.
- SVM training with the best parameters.

kernel	rbf	poly		
C	0.1	1	10	
gamma	1e-2	1e-4	scale	auto

Table 3.15. Grid search parameters for SVM.

- Evaluating the prediction results produced by the SVM on the test set using the following scores: accuracy, precision, recall, and f1.

- Picking the feature set that achieves the highest scores.

3.4.2 Images

Concerning the images, the extracted feature sets shown in the Table 3.14 must be evaluated.

For either *Base model 1* and *Base model 2*:

- I classify the test set images, obtaining baseline results. Each slide is made up of multiple crops and, therefore, the results are single-crop level. In a nutshell, for each crop belonging to a specific slide I have the healthy/tumor prediction.

- Considering the feature sets obtained after applying PCA, I train a Multi-Layer Perceptron for each of them.

- For each MLP obtained, I classify the test set, obtaining crop-level predictions within a whole slide.

- I aggregate the results from crop-level to slide-level, in two ways: a slide is globally classified as tumor if at least 10% of the crops composing it are classified as tumor. I also consider a majority voting approach, whereby a whole slide is classified as tumor if at least 50% of the crops it is composed of are classified as tumor.

- I analyze the results obtained and pick the feature set that classifies the test set in a way most closely resembling its baseline.

MLPs details

From Baseline model 1 and Baseline model 2, I get two different feature sets after applying PCA. This aspect means that I have to validate four different feature sets and thus train four MLPs. I describe the architecture details for each of them based on the input feature set.

Feature set 1: *Base model 1* + $PCA(n\_comp = 9)$

- Input layer: 9 nodes
- Hidden layers: $9 \rightarrow 6$ nodes
- Output layer: 2 nodes (softmax activation function)

Feature set 2: *Base model 1* + $PCA(n\_comp = 20)$

- Input layer: 20 nodes
- Hidden layers: $10 \rightarrow 6$ nodes
- Output layer: 2 nodes (softmax activation function)

Feature set 3: *Base model 2* + $PCA(n\_comp = 9)$

- Input layer: 9 nodes
- Hidden layers: $9 \rightarrow 6$ nodes
- Output layer: 2 nodes (softmax activation function)

Feature set 4: *Base model 2* + $PCA(n\_comp = 20)$

- Input layer: 20 nodes
- Hidden layers: $10 \rightarrow 6$ nodes
- Output layer: 2 nodes (softmax activation function)

The Table 3.16 summarizes the parameter choices for each feature set.

	Loss	Learning Rate	Optimizer	Epochs	Batch Size
Feature set 1	categorical cross-entropy	1e-4	Adam	50	32
Feature set 2	categorical cross-entropy	1e-4	Adam	40	32
Feature set 3	categorical cross-entropy	1e-4	Adam	30	32
Feature set 4	categorical cross-entropy	1e-4	Adam	40	32

Table 3.16. Parameters used for each MLP.

3.4.3 Image classification model

The validation of the extracted features allows to figure out the feature set that is most representative over its respective baseline in predicting the image test set. Since I will correlate these features with those extracted from the methylation dataset, the final image classification model is precisely the MLP that achieves the best results, meaning it predicts the test set similar to the respective baseline. To assess the similarity among the different predictions, I aggregate the results at the slide level and construct a confusion matrix, from which I then derive the accuracy, recall, precision, and f1-score. I compare the scores obtained from each MLP with the respective baseline and evaluate the most likely ones.

In making the decision, I evaluate an additional factor: the baseline results cover both tumor and healthy slides belonging to the image set test. In describing the data, I mentioned that I have access to more detailed information only for the tumor slides, that is, the percentage of tumor cells present in each slide. I decide to leverage this information to do a follow-up on the baseline that most closely matches the actual information in predicting the slides.

In short, to obtain the actual image classification model, I evaluate:

1. The baseline that most likely[5] predicts the tumor images of the test set concerning the available tumor percentage information.

[5]Whose percentage of crops classified as tumor is in a range [% tumor cells - 10, % tumor cells + 10]. Where "% tumor cells" is the available information about the percentage of tumor tissue in a given Whole slide.

2. The MLP that most likely classifies the entire test set as its respective baseline.

3.5 Integration method

At this point, I have at disposal the best feature set extracted from the methylation data and the best feature set extracted from the image data (for both the train and the test sets). The goal here is to understand whether or not there is some form of interdependency between the two types of data by exploring the correlation between them. To conduct such analysis, I exploit two distinct statistical methods: Pearson correlation and Mutual Information. I choose these two methods precisely to generate a more comprehensive analysis: the former is widely used for its simplicity but requires some assumptions about the distribution of data and measures linear relationships; the latter is undoubtedly more generic, free-distribution and able to measure nonlinear relationships, but its implementation is not trivial.

3.5.1 Correlation threshold based approach

I perform the correlation for both the train and the test sets. In detail, for each of the crops belonging to each whole slide, I have a vector of extracted features; even if one image is divided into multiple crops, it belongs to a single patient. Instead, in the methylation dataset, each patient is associated with a single sample, and therefore a single vector of extracted features is available. Therefore, the correlation is performed between all the crops of a specific patient with his/her methylation data. The crops belonging to the tumor images are correlated with the corresponding tumor methylation sample, similarly for the healthy data. Of course, the two vectors to be correlated have to be of the same size (Figure 3.7).

Correlation values are used to discard all those image crops with a correlation value below a certain threshold: in the case of Mutual Information, I choose a threshold value equal to 0; as for Pearson, I discard crops that have a correlation value below the first quartile (25%) of the maximum correlation value.

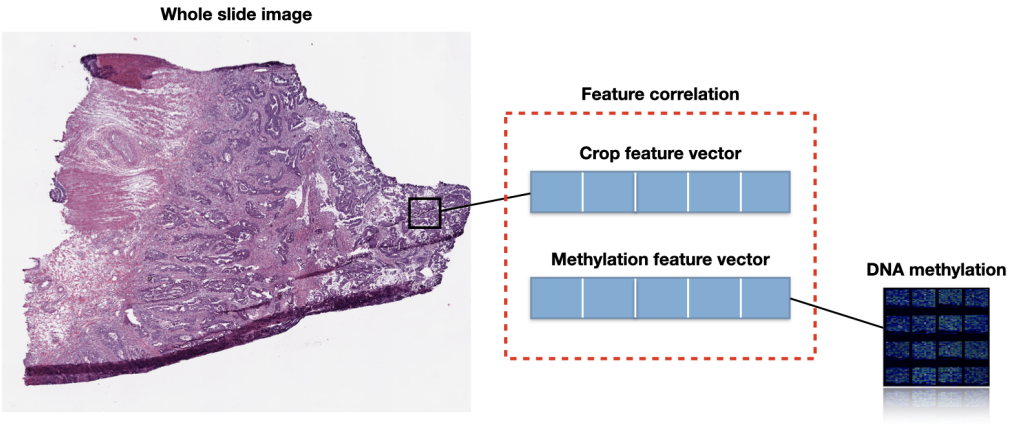

Figure 3.7. Correlation method. Each crop belonging to a given whole slide is correlated with the corresponding methylation value (derived from the same patient). The vectors are composed of the extracted features for both data types.

Correlation impact on image classification model

I compare the test set prediction results from the baseline image classification model with those obtained from the same predictive model, but without accounting for the under-threshold crops described above. Therefore, I obtain three different sets of prediction scores on the test set (Baseline case, MI case, and Pearson case). As discussed previously, classification results are crop-level. To be consistent, I aggregate all result sets from the crop-level to the slide-level in two ways: a slide is globally classified as tumor if at least 10% of its component crops are classified as tumor. I also consider a majority approach, whereby an entire slide is classified as tumor if at least 50% of its constituent crops are classified as tumor. I construct the usual confusion matrix for both aggregation strategies and calculate for each of the three cases accuracy, precision, recall, and f1-score. Observing these measures enables the assessment of which of the three cases achieves the best results.

Correlation and prediction heatmaps

To visualize the correlation values simultaneously, I reconstruct each image from the crops that compose it. Therefore, I generate a heatmap of correlation values, having a width and a length equal to the corresponding image to be reconstructed. The heatmap is helpful to observe, for each crop, the

correlation value associated with it and get a slide-level view of the correlation values distribution. Furthermore, I follow the same procedure for the classification results. This way, one can quickly check how the crops of a given slide have been classified and how many of them have been discarded because of the sub-threshold correlation value.

Pearson correlation details

In statistics, Pearson's correlation coefficient enables to measure a linear relationship between two sets of data.

Given two random variables X and Y, it is defined as the covariance of such variables, divided by the product of their standard deviation.

$$\rho = \frac{cov(X, Y)}{\sigma_x \sigma_y} \tag{3.3}$$

Where $cov(X, Y)$ is the covariance of X and Y, σ_x is the standard deviation of X and σ_y s the standard deviation of Y.

Pearson's correlation between the two vectors yields outcomes in a range [-1, 1]. Values around zero indicate weak or non-existing correlation, while values of -1 or +1 involve an exact linear relationship. Specifically, if two variables have a correlation equal to 1, they are positively correlated. Instead, if they have a correlation value equal to -1, they are negatively correlated (the relationship is inversely proportional). I exploit the Pearson correlation implementation provided by the SciPy library in its module named scipy.stats.pearsonr [67].

Mutual Information details

Mutual Information enables the measurement of nonlinear relationships between two random variables. As a concept closely related to entropy, it points out how much information can be gained from one random variable through observing another random variable. Specifically, it determines how different the joint distribution of the pair (X, Y) is from the product of the marginal distributions of X and Y.

Mathematically speaking, in terms of PDFs for continuous distributions:

$$I(X; Y) = \int_Y \int_X p_{(X,Y)}(x, y) \log \left(\frac{p_{(X,Y)}(x, y)}{p_X(x)\, p_Y(y)} \right) dx\, dy \tag{3.4}$$

Where $p_{(X,Y)}(x, y)$ is the joint probability density function, $p_X(x)$ and $p_Y(y)$ are respectively the marginal probability density functions of X and Y.

The main Mutual Information properties are:

- $I(X;Y) \geq 0$, that means it is a non-negative quantity.
- $I(X;Y) = I(Y;X)$, than indicates symmetry.
- If $I(X;Y) = 0$ means that X and Y are independent, because $p_{(X,Y)}(x, y) = p_X(x) \cdot p_Y(y)$

Hence, The MI method returns only non-negative values; results equal to 0 indicate that the feature vectors are independent of each other. The scikit-learn library provides an implementation of the method (the module name is sklearn.feature_selection.mutual_info_regression), which enables estimating mutual information for a continuous target variable. As described in the documentation, the implementation is based on nonparametric methods relying on entropy estimation from k-nearest neighbors distances [68] [69]. Both methods are grounded on the original idea first proposed in [70]. Considering the methylation feature vector as the target vector, the method checks how much dependence exists between the image feature vector and such target vector.
In the implementation provided by scikit-learn, it is required to choose an appropriate value for the parameter k, representing the number of neighbors to use for MI estimation for continuous variables. Higher values reduce the variance of the estimation but might introduce a bias. The default value proposed by the library is 3.

Therefore, I optimize the parameter search on the train set by evaluating a value of k equal to 1, 3, and 5. For each value of k, the flow is as follows:

- Perform feature correlation.
- Following the approach described previously, classify the train set using the image classification model, discarding the crops of a given slide whose correlation value is below the chosen threshold, that is 0.
- Evaluate aggregate classification results to choose the best k, considering both the 10% based and majority voting based global classification of a whole slide.

The Tables 3.17 and 3.18 reveal, for each evaluated value of k, respectively, the train set scores considering both the 10% based and majority voting based global classification of a whole slide. From the results, I infer that the best k to use is 5.

	accuracy	**precision**	**recall**	**f1**
MI (k=1)	0.860465	1.0	0.781818	0.877551
MI (k=3)	0.930233	1.0	0.877551	0.934783
MI (k=5)	0.965116	1.0	0.934783	0.966292

Table 3.17. Accuracy, precision, recall and f1 score for each value of k, 10% based.

	accuracy	**precision**	**recall**	**f1**
MI (k=1)	1.0	1.0	1.0	1.0
MI (k=3)	1.0	1.0	1.0	1.0
MI (k=5)	1.0	1.0	1.0	1.0

Table 3.18. Accuracy, precision, recall and f1 score for each value of k, majority voting based.

Chapter 4

Results

In this chapter, I present all the results obtained. Starting from validating the feature sets resulting from the extraction phase, I illustrate all the classification results for methylation and image data. Then, using the prediction results of the model selected to classify the images, I point out how the integrative method described in the previous chapter impacts these results.

4.1 Methylation: extracted feature sets

As reported in Table 3.9 and Table 3.11, I evaluate multiple feature sets extracted from methylation data, exploiting two different approaches, one based on PCA and the other based on Autoencoder. Table 4.1 shows, for each feature set, the best parameters, assessed with cross-validation, used to train the Support Vector Machine. Table 4.2 displays the scores derived from the methylation test set classification for each feature set. The highest scores are marked in grey.

	SVM_kernel	SVM_C	SVM_gamma
PCA (0.65)	poly	10	0.0001
PCA (0.80)	rbf	1	scale
PCA (0.99)	rbf	1	scale
Encoder 1	poly	0.1	auto
Encoder 2	poly	0.1	0.01
Encoder 3	poly	0.1	0.01
Encoder 4	poly	0.1	auto

Table 4.1. Best SVM parameters for each methylation extracted feature set.

	# feat. extr.	accuracy	precision	recall	f1
PCA (0.65)	9	0.993289	1	0.991597	0.995781
PCA (0.80)	23	0.986577	1	0.983193	0.991525
PCA (0.99)	76	0.986577	1	0.983193	0.991525
Encoder 1	20	0.973154	1	0.966387	0.982906
Encoder 2	150	0.986577	1	0.983193	0.991525
Encoder 3	512	0.979866	1	0.97479	0.987234
Encoder 4	20	0.986577	0.983471	1	0.991667

Table 4.2. SVM classification scores for each methylation extracted feature set.

4.2 Images: extracted feature sets

The organization is as so: for both *Base model 1* and *Base model 2*, I report the histograms of the test set classification results, keeping the tumor slides separate from the healthy slides. Next, just for the tumor slides, I display in a bar chart how the percentage of tumor cells, healthy and belonging to another tissue (the stromal one), is distributed based on the information collected by the GDC Data Portal (the ground-truth). Then, I compare that information with the classification results of the two models mentioned above. Lastly, for all MLPs trained on the extracted feature sets, I show the classification results aggregated on slide-level via tables. Then, as described in the previous chapter, I present both results using aggregation 10% based and majority voting based.

4.2.1 Base model 1

Figure 4.1 presents the classification results for tumor slides belonging to the test set. Whereas Figure 4.2 shows the results for healthy slides.

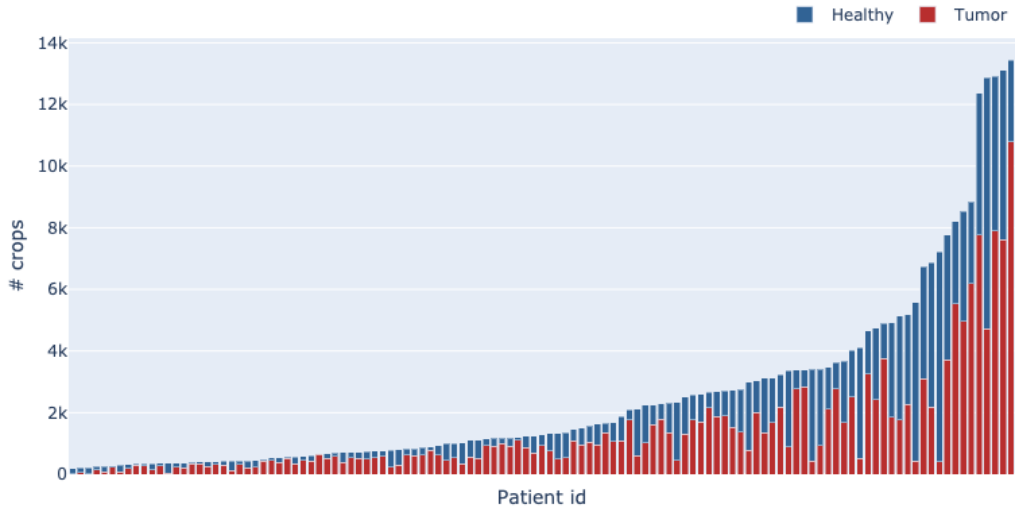

Figure 4.1. Base model 1 classification results for tumor slides. For each patient (each WSI), the stacked bar plot reveals the number of crops classified as tumor (in red) and the number of crops classified as healthy (in blue).

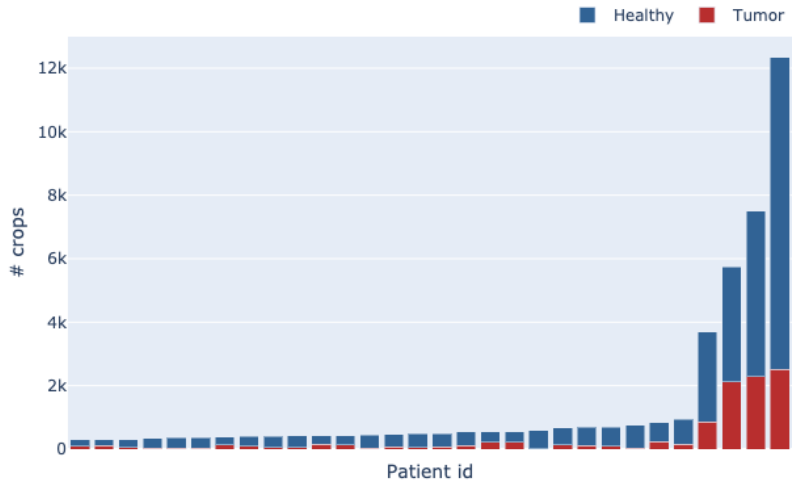

Figure 4.2. Base model 1 classification results for healthy slides. For each patient (each WSI), the stacked bar plot reveals the number of crops classified as tumor (in red) and the number of crops classified as healthy (in blue).

4.2.2 Base model 2

Figure 4.3 presents the classification results for tumor slides belonging to the test set. Whereas Figure 4.4 shows the results for healthy slides.

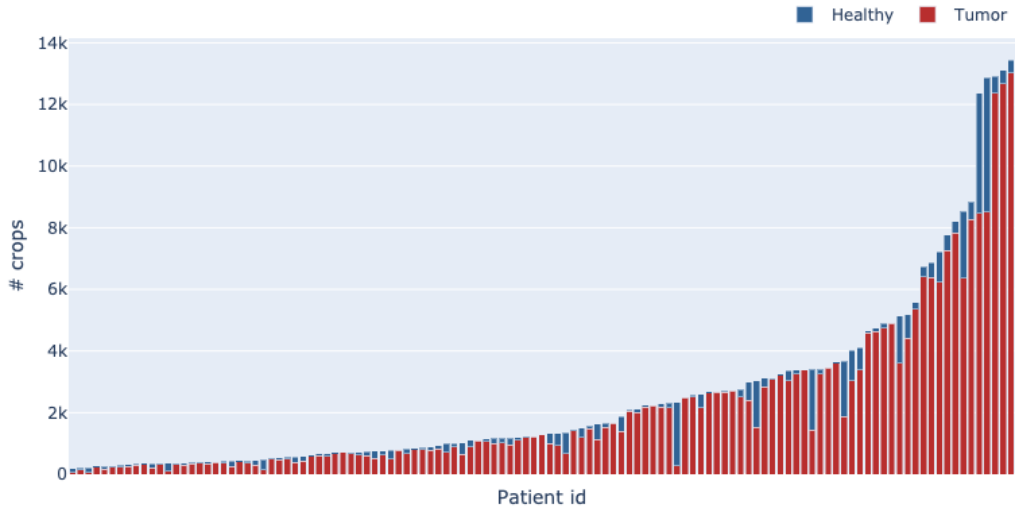

Figure 4.3. Base model 2 classification results for tumor slides. For each patient (each WSI), the stacked bar plot reveals the number of crops classified as tumor (in red) and the number of crops classified as healthy (in blue).

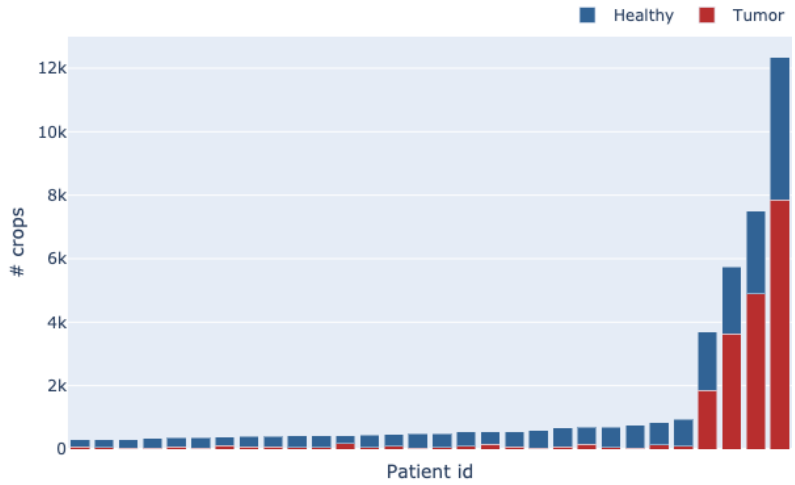

Figure 4.4. Base model 2 classification results for healthy slides. For each patient (each WSI), the stacked bar plot reveals the number of crops classified as tumor (in red) and the number of crops classified as healthy (in blue).

4.2.3 Tumor slides groundtruth

The Figure 4.5 indicates the percentage of tumor cells, healthy cells, or cells belonging to stromal tissue in each whole slide for tumor slides. This information is considered a ground truth (GTH) more detailed than the global label associated with the entire slide I have available.

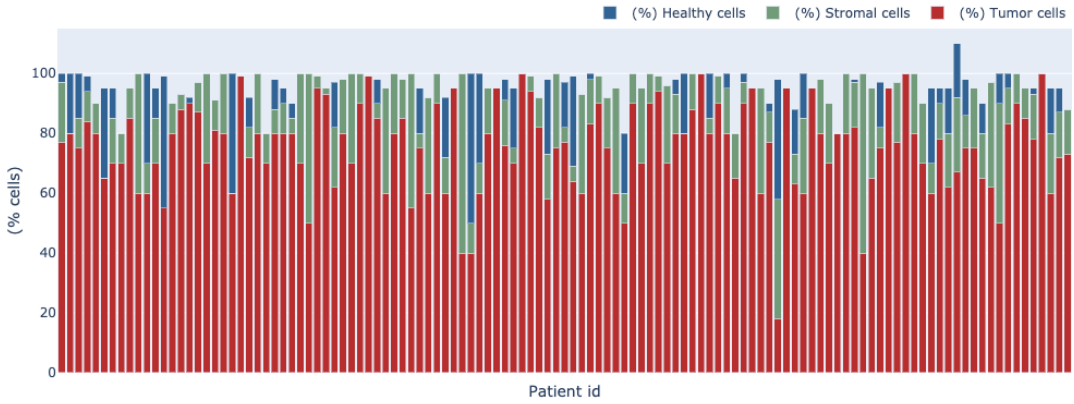

Figure 4.5. Tumor slides. For each patient (each slide) the percentage of tumor (red color), healthy (blue color) and stromal cells (green color) as provided by the GDC Data Portal is shown. This information is considered as a groundtruth.

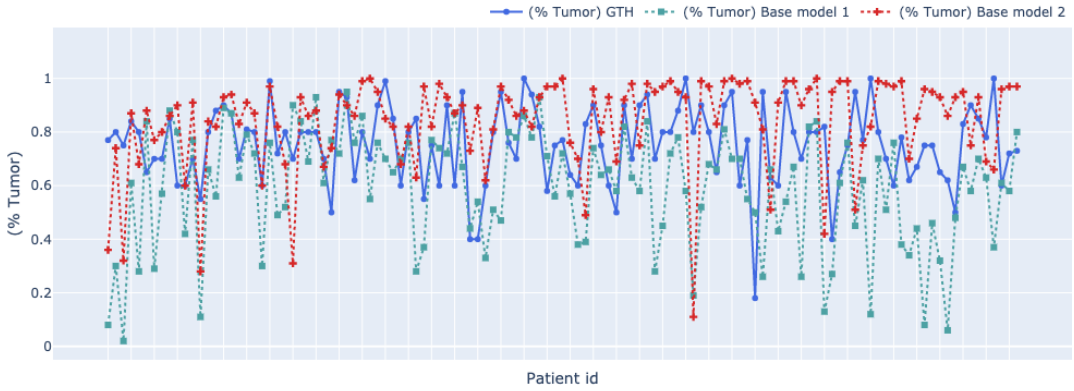

Figure 4.6. Tumor slides. Percentage of crops classified as tumor for each patient (each slide) by using Base model 1 (green color) and Base model 2 (red color), in relation to groundtruth information (blue color).

In Figure 4.6, it is visible how Base model 1 and Base model 2 classify tumor slides, compared to ground truth. In detail, for each slide, the percentage of crops classified as tumor is reported concerning the percentage of tumor cells provided by ground truth. To provide a qualitative description of the obtained results, I assume that a prediction is comparable to the ground truth if it is included in a range [% tumor cells$_{GTH}$ - 10, % tumor cells$_{GTH}$ + 10]. Therefore, it is inferred that Base model 1 classifies 37/119 slides in a comparable way to the ground truth information, while the number of slides for Base model 2 increases to 44/119.

4.2.4 Comparison results

The resulting tables allow a compact assessment of how significantly the predictions of the feature-trained models deviate from the corresponding baseline models. The confusion matrix is the tool that most allows observing in a straightforward way how each model behaves. As anticipated in the previous chapter (Figure 3.4), TP (True Positives) and TN (True Negatives) are the amount of tumor and healthy samples that are correctly predicted, respectively. Similarly, FP (False Positives) and FN (False Negatives) are the amount of tumor and healthy samples that are mispredicted, respectively.

The Tables 4.3 and 4.4 report the classification results in terms of confusion matrices, while the Tables 4.5 and 4.6 output all the scores derived from such matrices.

		Tumor slides (119)		**Healthy slides (30)**	
	# feat. extr.	**TP**	**FN**	**TN**	**FP**
Base model 1	None	115	4	6	24
MLP - Feature set 1	9	119	0	0	30
MLP - Feature set 2	20	119	0	0	30
Base model 2	None	119	0	9	21
MLP - Feature set 3	9	118	1	9	21
MLP - Feature set 4	20	118	1	11	19

Table 4.3. Confusion matrix (flattened in a row) for Base models and MLPs trained on extracted feature sets, assuming 10% threshold.

		Tumor slides (119)		Healthy slides (30)	
	# feat. extr.	TP	FN	TN	FP
Base model 1	None	83	36	30	0
MLP - Feature set 1	9	105	14	12	18
MLP - Feature set 2	20	109	10	26	4
Base model 2	None	112	7	27	3
MLP - Feature set 3	9	113	6	27	3
MLP - Feature set 4	20	110	9	27	3

Table 4.4. Confusion matrix (flattened in a row) for Base models and MLPs trained on extracted feature sets, assuming 50% threshold (majority voting).

	# feat. extr.	accuracy	precision	recall	f1
Base model 1	None	0.812081	0.966387	0.827338	0.891473
MLP - Feature set 1	9	0.798658	1.0	0.798658	0.88806
MLP - Feature set 2	20	0.798658	1.0	0.798658	0.88806
Base model 2	None	0.85906	1.0	0.85	0.918919
MLP - Feature set 3	9	0.852349	0.991597	0.848921	0.914729
MLP - Feature set 4	20	0.865772	0.991597	0.861314	0.921875

Table 4.5. Prediction scores for Base models and MLPs trained on extracted feature sets, assuming 10% threshold.

	# feat. extr.	accuracy	precision	recall	f1
Base model 1	None	0.758389	0.697479	1.0	0.821782
MLP - Feature set 1	9	0.785235	0.882353	0.853659	0.867769
MLP - Feature set 2	20	0.90604	0.915966	0.964602	0.939655
Base model 2	None	0.932886	0.941176	0.973913	0.957265
MLP - Feature set 3	9	0.939597	0.94958	0.974138	0.961702
MLP - Feature set 4	20	0.919463	0.92437	0.973451	0.9482

Table 4.6. Prediction scores for Base models and MLPs trained on extracted feature sets, assuming 50% threshold (majority voting).

4.2.5 Final image classification model

The selected image classification model is derived from *Base model 2*. Specifically, I decide to use the MLP trained on Feature set 3. Histograms in Figure 4.7 and Figure 4.8 bring out the classification results of the selected model for tumor and healthy slides at crop level.

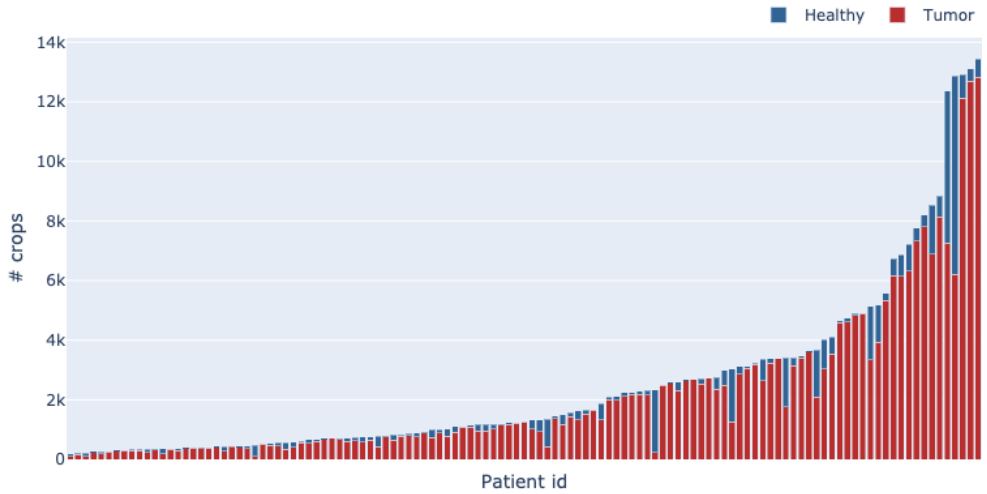

Figure 4.7. Baseline classification results for tumor slides. For each patient (each WSI), the stacked bar plot reveals the number of crops classified as tumor (in red) and the number of crops classified as healthy (in blue).

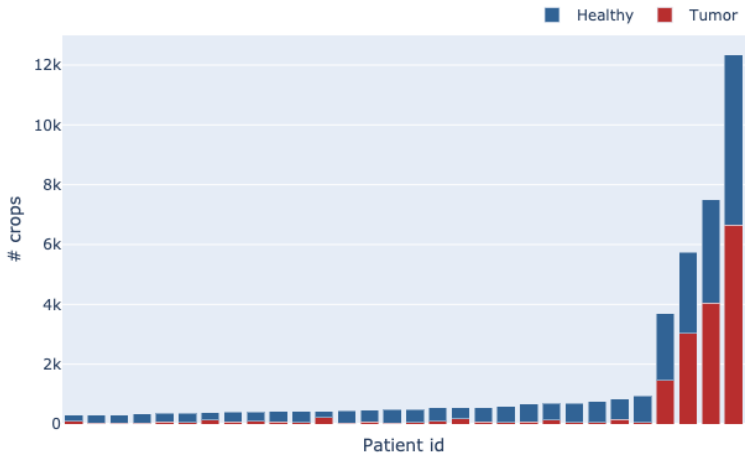

Figure 4.8. Baseline classification results for healthy slides. For each patient (each WSI), the stacked bar plot reveals the number of crops classified as tumor (in red) and the number of crops classified as healthy (in blue).

4.3 Integration method: comparison results

The comparative results reveal how the image classification model behaves after discarding those sub-threshold correlation crops. Specifically, after correlating each crop of each slide with its respective methylation sample, I classify test set slides in three different ways:

1. Not discarding any crops (Baseline case).
2. Removing crops below Pearson's correlation threshold (Pearson case).
3. Removing crops below MI's correlation threshold (MI case).

The Tables 4.7 and 4.8 report the classification results in terms of confusion matrices, while the Tables 4.9 and 4.10 output all the scores derived from such matrices.

	Tumor slides (119)		**Normal slides (30)**	
	TP	**FN**	**TN**	**FP**
Baseline	118	1	9	21
Pearson	118	1	12	18
MI	118	1	20	10

Table 4.7. Confusion matrix (flattened in a row) for Baseline, Pearson and MI case, assuming 10% threshold.

	Tumor slides (119)		**Normal slides (30)**	
	TP	**FN**	**TN**	**FP**
Baseline	113	6	27	3
Pearson	100	19	13	17
MI	101	18	30	0

Table 4.8. Confusion matrix (flattened in a row) for Baseline, Pearson and MI case, assuming 50% threshold (majority voting).

	accuracy	precision	recall	f1
Baseline	0.852349	0.991597	0.848921	0.914729
Pearson	0.872483	0.991597	0.867647	0.92549
MI	0.926174	0.991597	0.921875	0.955466

Table 4.9. Prediction scores for Baseline, Pearson and MI case, assuming 10% threshold.

	accuracy	precision	recall	f1
Baseline	0.939597	0.94958	0.974138	0.961702
Pearson	0.758389	0.840336	0.854701	0.847458
MI	0.879195	0.848739	1.0	0.918182

Table 4.10. Prediction scores for Baseline, Pearson and MI case, assuming 50% threshold (majority voting).

Chapter 5

Discussion

This chapter focuses on interpreting the results presented previously by motivating all decisions that have been taken.

5.1 Methylation: final extracted feature set

As demonstrated in the preliminary part of the method, methylation is a data type that can address well the classification task among the two classes (tumor, healthy). As a matter of fact, in Table 3.7, I emphasize that it is possible to train four different genomic classifiers and achieve very high levels of accuracy with each of them. It should be argued that methylation data is not the only genomic source available. However, after carefully evaluating other omic data types (e.g., transcriptomics), I chose methylation data due to the reliability of the obtained models. In addition, although methylation and histological images refer to the same patient and tumor, DNA methylation data could not refer precisely to the same piece of tissue used to get the histological image but to a very close area.

Regarding the feature sets extracted by methylation, I note from Table 4.2 that the PCA-based approach with explained variance equal to 65% is the best. The SVM trained on this feature set achieves better accuracy, precision, and f1 scores than the other evaluated approaches. It is also challenging to work on such a small number of features, which is representative of the dataset; this number is the same used to reduce the dataset before training the genomic classifiers. From the same table, I note that the feature extraction approach through Encoder 4 allows reaching the maximum level of recall, going to extract 20 features. It is the initial reason that leads me to

evaluate for the image data a set of extracted features equal to 9 and equal to 20 to perform the integration process with two equal-sized vectors.

5.2 Images: final extracted feature set

The image data is undoubtedly more challenging to handle than the methylation one. The slides collected from the TCGA repository were revealed to be highly noisy, with a presence of tissue types not referable to the only two classes healthy or tumor. Figure 4.5 emphasizes how for every tumor slide, there is often a non-negligible amount of stromal tissue present. Unfortunately, this information is not known for healthy slides. This issue has a substantial impact on the performance of the convolutional network developed for feature extraction since I train a classifier on two classes, not having the correct labeling of each crop, but only the global label of the whole slide.

The first point to consider to evaluate the best set of features extracted from the images is how *Base model 1* and *Base model 2* behave in the classification task. Regarding tumor slides, it is evident (Figures 4.1 and 4.3) that Base model 2 has a tendency to classify more crops as tumor than Base model 1. It is justified by the fact that Base model 1 is not trained on the TCGA train set images but has learned to recognize only the information provided by the external ROIs set. For the healthy slides (Figures 4.2 and 4.4) the two models behave in a rather similar way, even if Base model 1 tends to classify more crops as tumor in the slides of smaller size. As for the slides of larger size, it is evident from both models (more so in Base model 2) that a large part of the crops is not considered healthy, but this is right because those slides are extremely noisy and contain a significant amount of tissue that is neither healthy nor tumor.
From Figure 4.6, I note that the two models behave quite similarly to the ground truth information in classifying tumor slides. However, as anticipated, Base model 2 proves to be slightly more likely to agree with the ground truth information. To select the set of extracted features to be used in the integrative approach, I check how far the predictions of the trained models on such extracted features deviate from the respective Base models. The confusion matrices shown in the Table 4.3 and in the Table 4.4 clearly demonstrate how Feature set 3 (obtained from the Base model 2 + PCA(n_comp=9) pipeline) turns out to be the most representative of the

respective Base model. In fact, with the 10%-based global slide prediction approach, it behaves exactly like Base model 2, except for one tumor slide. The same holds by considering the majority voting approach. The tables showing the scores (Table 4.5 and Table 4.6) bring a further support to the observation just given. I consequently decide to select Feature set 3 since:

1. Base model 2 used to obtain this feature set predicts tumor slides more accurately than Base model 1 compared to ground truth data.

2. It is the most representative feature set in comparison to its Base model.

3. It consists in getting, for each crop of each whole slide, a number of features equal to 9, that is, the same amount of features extracted from the methylation dataset for each sample.

The histograms in the Figures 4.7 and 4.8 show the classification results of what I consider the actual image classification baseline model, meaning the Multi-Layer Perceptron trained on the selected Feature set 3. Observing the bar charts, one can notice that they are almost identical to the classification results presented previously (Figure 4.3 and 4.4), which concern the Base model 2.

5.3 Integration method

From the analysis of the results described above, it is inferred that to perform the correlation task, each methylation sample is characterized by a number of features extracted equal to 9 and each crop from each slide image. The classification results for each slide are reported in an aggregate form to decide whether a slide is globally classified as tumor or not, based on the number of crops classified as tumor. Assuming a 10% crop threshold, as shown in the Table 4.7, for tumor slides, the 3 cases analyzed behave in the same way, classifying 118/119 tumor slides (TP). The MI method differs in healthy slides, increasing the number of TNs compared to Baseline. In the Pearson case, there is a slight increase in TNs, but in a smaller form. This finding indicates that the Mutual Information-based method is more successful in identifying crops classified as tumor in the healthy slide case, thus allowing them to be eliminated and improving performance. This result is confirmed in the case of the global classification of the slide through majority voting, as shown in Table 4.8. However, in such a table, it can be noted that, the performance decreases using the Pearson and MI methods, compared to the

Baseline considering the majority voting approach. This point indicates that in many cases, crops that the model classifies as tumor are discarded. The tables (Table 4.9 and 4.10) showing the scores derived from the confusion matrices confirm what has to be stated. Based on the results, I conclude by saying that assuming the 10% threshold, the MI based method guarantees better performance, whereas assuming the 50% threshold, the Baseline achieves the highest scores in the classification task.

5.3.1 Main issues and possible improvement solutions

Using the majority voting approach allows one to understand if there is an excessive number of crops in a certain slide that is inconsistently classified concerning the original global label. However, the results are presented in aggregate form, based on the assumption that the Baseline model can identify accurately whether a certain crop is healthy or tumor. As anticipated above, this is the real challenge. In the supplementary material at the end of the work, I show for some interesting example cases how the image classification model behaves by reconstructing each slide through the heatmaps. I realize that exploiting the correlation between the image and the methylation, crop by crop, has positive feedback because, in many cases, it leads to identifying the areas of the slide where there is tumor tissue (or not). In addition, where the removal technique fails to identify the background, I find that the approach employed succeeds in identifying such background areas and consequently removing them. However, having such noisy slides, with tissues differing from healthy or tumor one, undermines the image feature extractor model (and therefore also the image classification model) in some cases.
The main issues I notice through the heatmaps are:

- The model associates the other tissue types (predominantly stromal cells) either with the healthy class or the tumor class without a well-identified criterion. This aspect influences the classification results quite dramatically.

- As a ripple effect of the prior issue, the correlation succeeds in identifying areas of different tissue from each other, however correlating positively with the wrong areas of tissue.

Possible strategies for improvement are:

- Collect images to learn a three-case problem, providing material to make the model learn other tissue besides healthy and tumor ones.
- Consider an approach that modifies classification results not only on the discard of sub-threshold correlation crops. Specifically, the correlation value might be used to give more or less weight to the predictions made by the model.

Chapter 6

Conclusions

This work stands within an emerging interdisciplinary research field, discussing an approach for integrating two types of biomedical data: histological images and DNA methylation. The latter can reveal transcriptional regulation mechanisms and, consequently, it is suitable to study pathological conditions, particularly cancer.
The objective is to train an image classification model to predict malignancy in histological images and investigate how methylation data affects prediction performance by exploiting the correlation between features extracted from the two data types. To validate the approach, I use colon cancer patient data (methylation samples and whole slide images) from TCGA repository and ROIs from a previous case study.
Key steps are:

- Preprocessing of both data types. At this stage, the images are cut into smaller sized crops, discarding the background.
- Feature extraction, evaluating the extraction of different feature sets for both data types, exploiting various ML/DL based approaches.
- Feature validation. The extracted feature sets are used to train genomic and image classifiers separately. For each of the two data types, the classifier that achieves better performance in predicting the test set is chosen, and, consequently, the best set of extracted features is elected. The selected image classification model becomes the Baseline.
- Feature correlation between the best set of extracted features for images and the best set of extracted features for methylation, by employing two statistical methods: Pearson correlation and Mutual Information.

Specifically, each crop of each whole slide of a given patient is correlated with the respective methylation sample.

- Implementation of an approach that modifies the image classification results (Baseline case) based on the crop correlation value: the crop is discarded if its correlation value is below the first quartile (25%) of the maximum correlation value (Pearson case) or equal to 0 (MI case).
- Analysis of the prediction results obtained for all three cases discussed aforementioned. Considering that, for each slide, the classification provides crop-level results, I consider two aggregation strategies:
  1. A whole slide is globally labeled as tumor if at least 10% of the crops are labeled as tumor.
  2. A whole slide is globally labeled as tumor if at least 50% of the crops are labeled as tumor (majority voting).
- Whole slide image reconstruction using correlation and prediction heatmaps to observe, for each crop, the associated correlation value and prediction provided by the image classification model, respectively.

From the results analysis, I derive that:

- Assuming a 10% based threshold, the MI based method guarantees better performance in classifying healthy slides, while the tumor slide classification is equal for the three cases. Therefore, considering both classes, the MI-based method guarantees better performances.
- Assuming a majority voting based threshold, the Baseline achieves the highest scores in the classification task.

I gather from the slide reconstruction that:

- Exploiting the correlation between the image and the methylation, crop by crop, has positive feedback because in many cases leads to identify the areas of the slide where there is tumor tissue (or not).
- TCGA images contain tissue that cannot be associated with the healthy class or tumor class. The model assigns the other tissue types present (predominantly stroma) to either the healthy or tumor class without a well-identified criterion. This point affects the classification results quite dramatically. As a ripple effect of this issue, the correlation succeeds in identifying areas of different tissue from each other, however correlating positively with the wrong areas of tissue.

The main challenges of this analysis mainly derive from the image data coming from the TGCA repository. Although the database provides a global label for each image, there is often a non-negligible percentage of other tissues inside (e.g., stromal tissue), which adds noise and introduces an error in the training of the models. In detail, it would be necessary to have at least a third-class available to distinguish between healthy, tumor, and other tissue types to improve the reliability of the results. In addition, it could improve the performance of the feature extractor model and, consequently, the correlation values.

Appendix A

Example slides

This appendix is meant to illustrate some example slides, by showing for each of them the original image and classification results considering Baseline, Pearson and MI case. I also display the correlation and prediction heatmaps. In correlation heatmaps, the higher the correlation of the crop the lighter its color will be. In classification heatmaps, each crop is colored red if it has been predicted as tumor, blue if the prediction is healthy. Getting a visual of the reconstructed image permits an immediate view of how the approach works, which is definitely useful to observe crop by crop the correlation values and, clearly, also the predictions of the image classification model. In the original images can be observed the presence of areas of tissue that cannot be associated with either the healthy or tumor class. This phenomenon is particularly impactful in the prediction of healthy slides, as the presence of unknown tissue confuses the classifier and leads to incorrect predictions.

A.1 Tumor slide 1

The slide has a dirty background, and the removal approach did not completely remove it, as can be seen in the Figure A.2 (left). The example shows how both methods detects dirty background areas and removes them. In this specific case, Pearson method succeeds in cleaning the slide better.

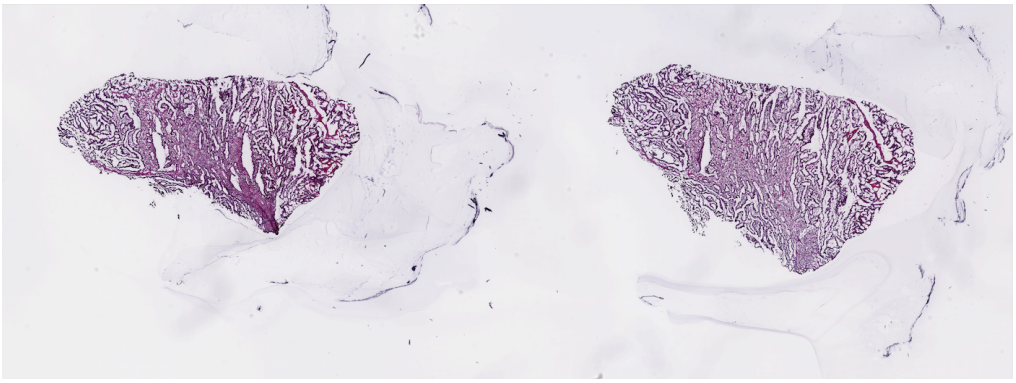

Figure A.1. Original image.

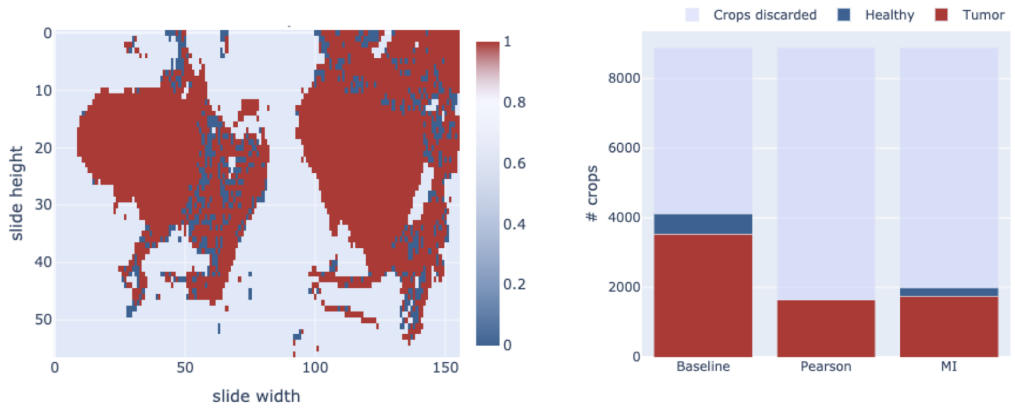

Figure A.2. On the left, the heatmap shows how the Baseline model classifies each crop on the slide. On the right is a stacked bar chart showing how the predictions vary in the Baseline, Pearson, and MI case.

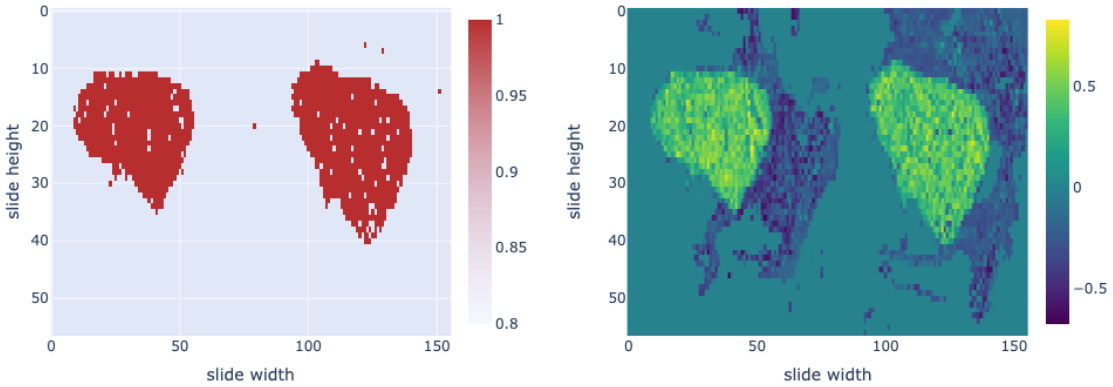

Figure A.3. Pearson case. On the left, the heatmap illustrates how the model classifies each crop on the slide, not considering sub-threshold correlation crops. On the right is shown the correlation value heatmap.

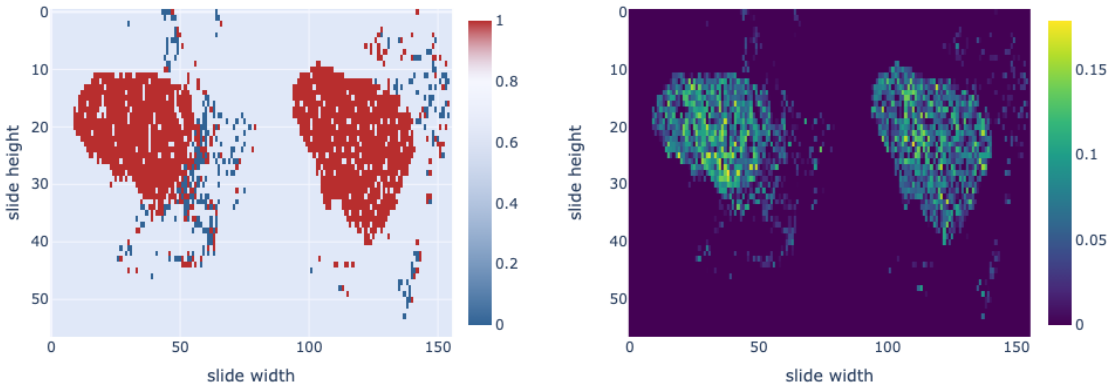

Figure A.4. MI case. On the left, the heatmap illustrates how the model classifies each crop on the slide, not considering sub-threshold correlation crops. On the right is shown the correlation value heatmap.

A.2 Tumor slide 2

This slide shows one of the main issues encountered: Pearson based approach identifies a correlation with the methylation sample, but comes out negative. For this reason in the final analysis almost all the crops are thrown away, penalizing such approach compared to the one based on MI.

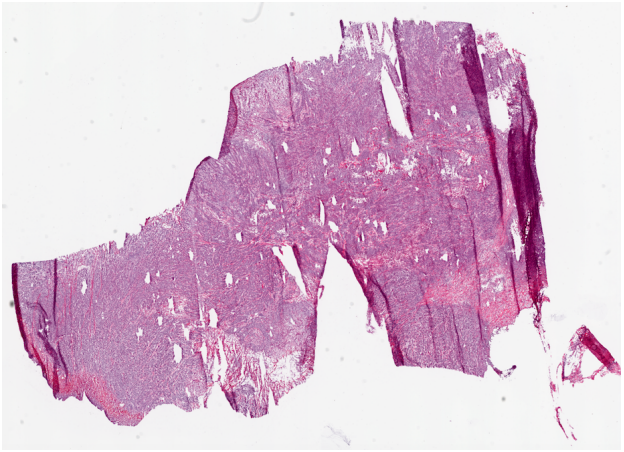

Figure A.5. Original image.

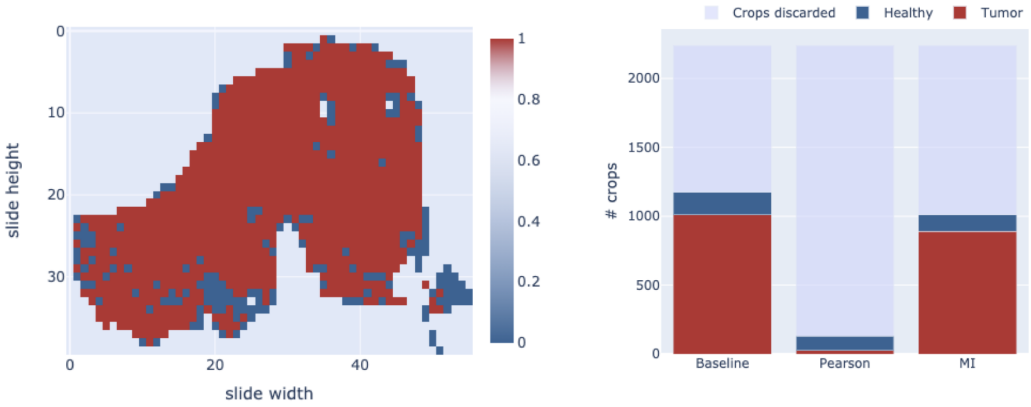

Figure A.6. On the left, the heatmap shows how the Baseline model classifies each crop on the slide. On the right is a stacked bar chart showing how the predictions vary in the Baseline, Pearson, and MI case.

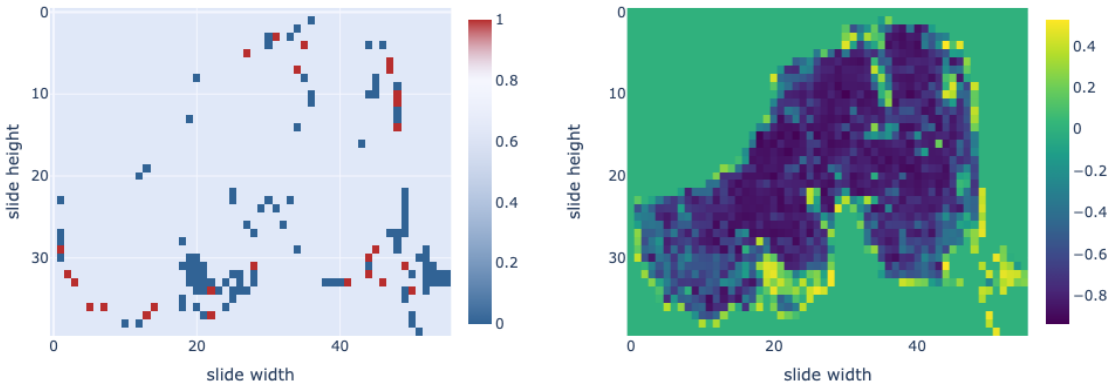

Figure A.7. Pearson case. On the left, the heatmap illustrates how the model classifies each crop on the slide, not considering sub-threshold correlation crops. On the right is shown the correlation value heatmap.

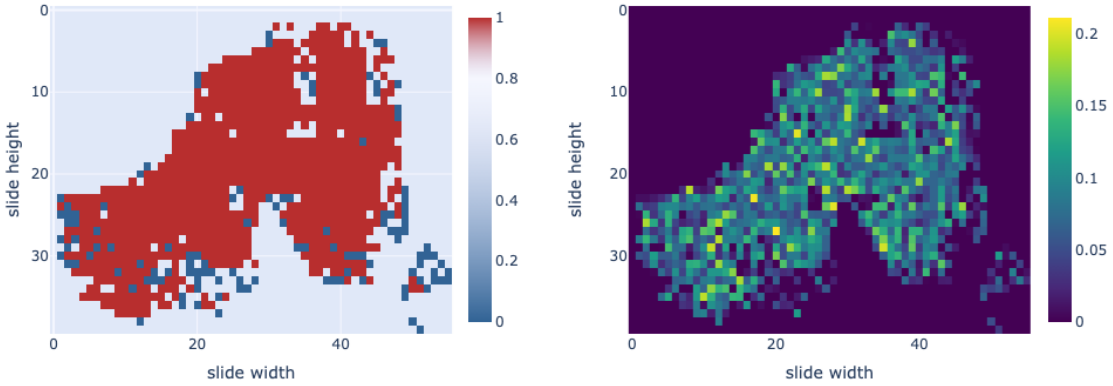

Figure A.8. MI case. On the left, the heatmap illustrates how the model classifies each crop on the slide, not considering sub-threshold correlation crops. On the right is shown the correlation value heatmap.

A.3 Tumor slide 3

In the slide, it is noted that the Pearson-based approach is more successful in cleaning up the slide of values predicted as healthy than the MI-based approach.

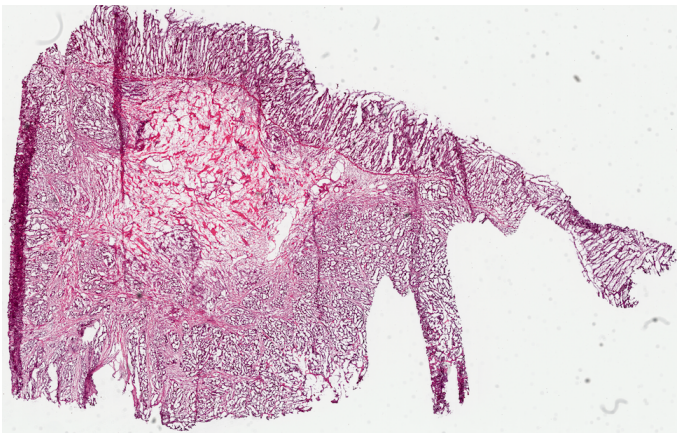

Figure A.9. Original image.

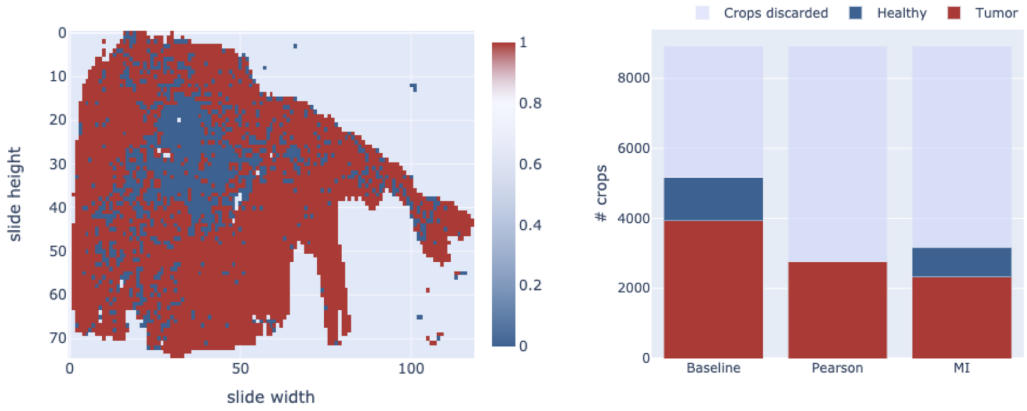

Figure A.10. On the left, the heatmap shows how the Baseline model classifies each crop on the slide. On the right is a stacked bar chart showing how the predictions vary in the Baseline, Pearson, and MI case.

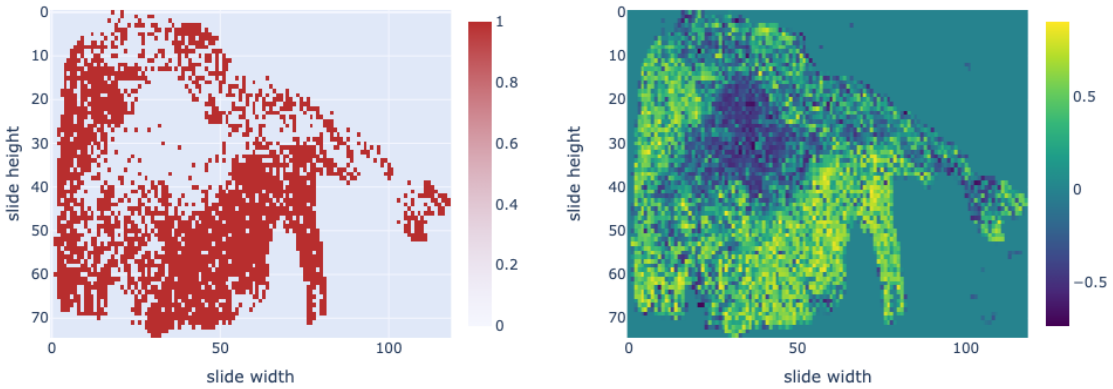

Figure A.11. Pearson case. On the left, the heatmap illustrates how the model classifies each crop on the slide, not considering sub-threshold correlation crops. On the right is shown the correlation value heatmap.

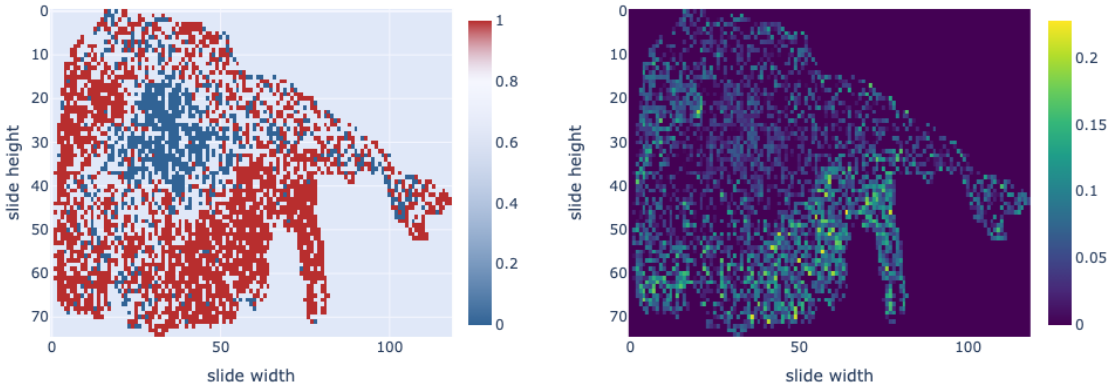

Figure A.12. MI case. On the left, the heatmap illustrates how the model classifies each crop on the slide, not considering sub-threshold correlation crops. On the right is shown the correlation value heatmap.

A.4 Healthy slide 1

The slide visibly shows a presence of different tissues (the healthy one is on the left). The MI-based correlation heatmap exhibits this diversity, whereas Pearson strongly negatively correlates also the healthy area.

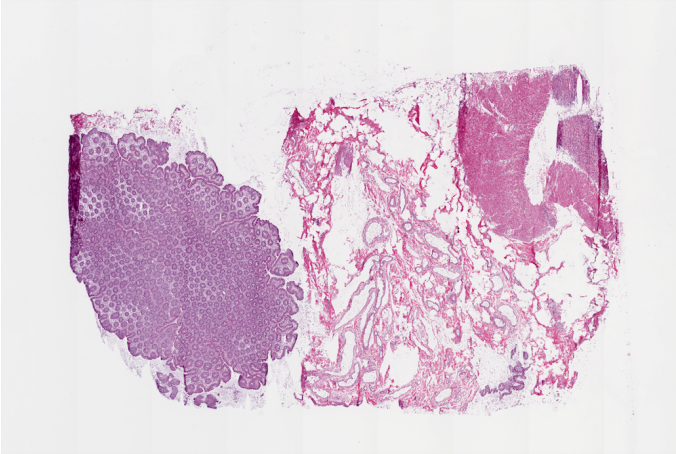

Figure A.13. Original image.

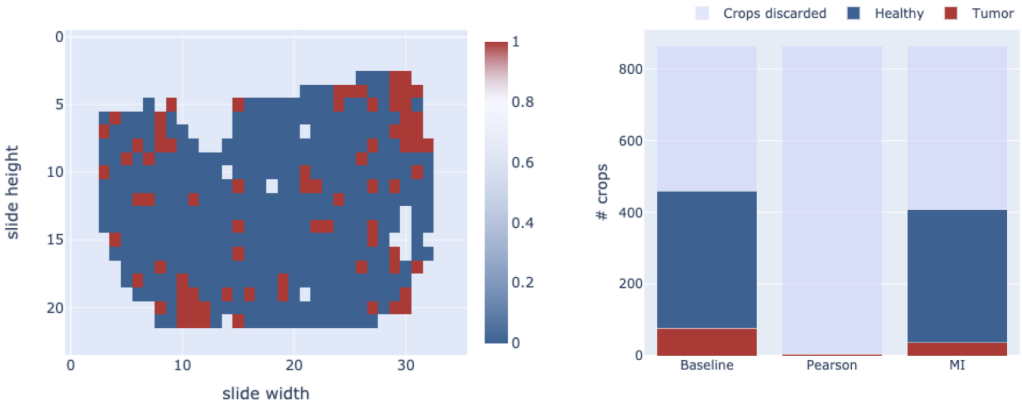

Figure A.14. On the left, the heatmap shows how the Baseline model classifies each crop on the slide. On the right is a stacked bar chart showing how the predictions vary in the Baseline, Pearson, and MI case.

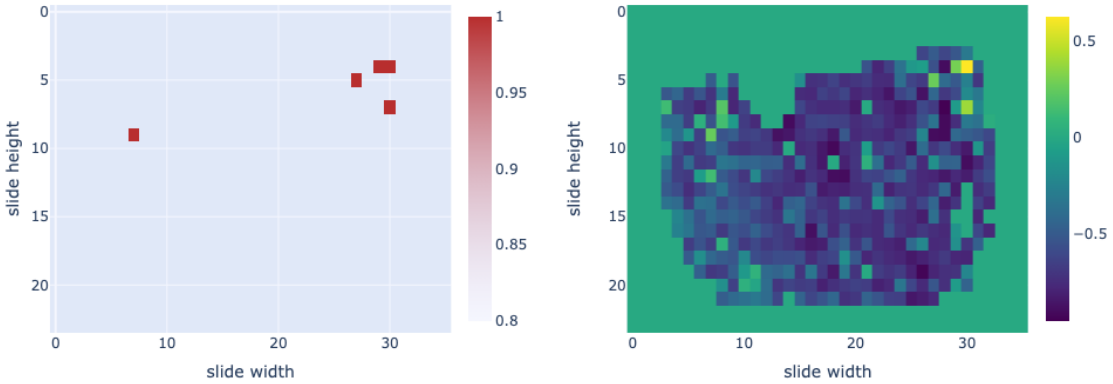

Figure A.15. Pearson case. On the left, the heatmap illustrates how the model classifies each crop on the slide, not considering sub-threshold correlation crops. On the right is shown the correlation value heatmap.

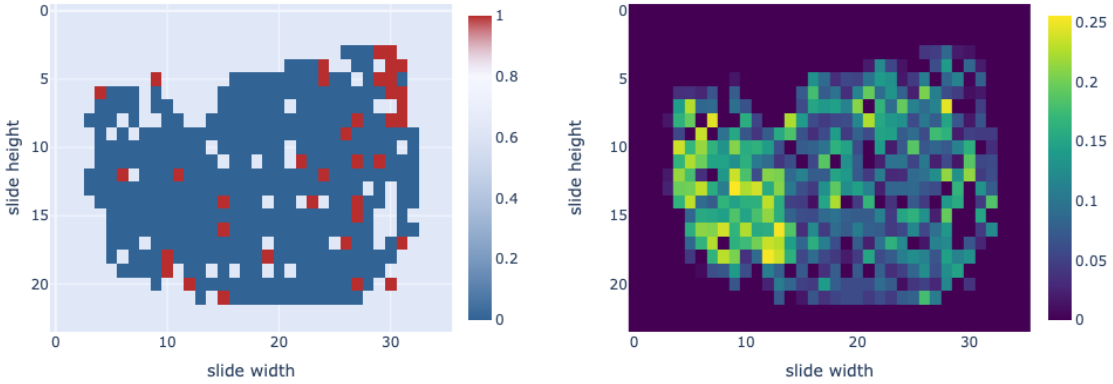

Figure A.16. MI case. On the left, the heatmap illustrates how the model classifies each crop on the slide, not considering sub-threshold correlation crops. On the right is shown the correlation value heatmap.

A.5 Healthy slide 2

This example shows a very noisy slide, making it difficult to classify the different tissue zones correctly. Correlation values are also affected.

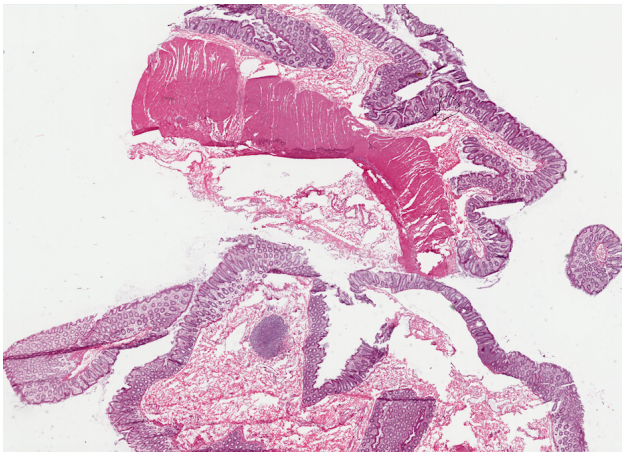

Figure A.17. Original image.

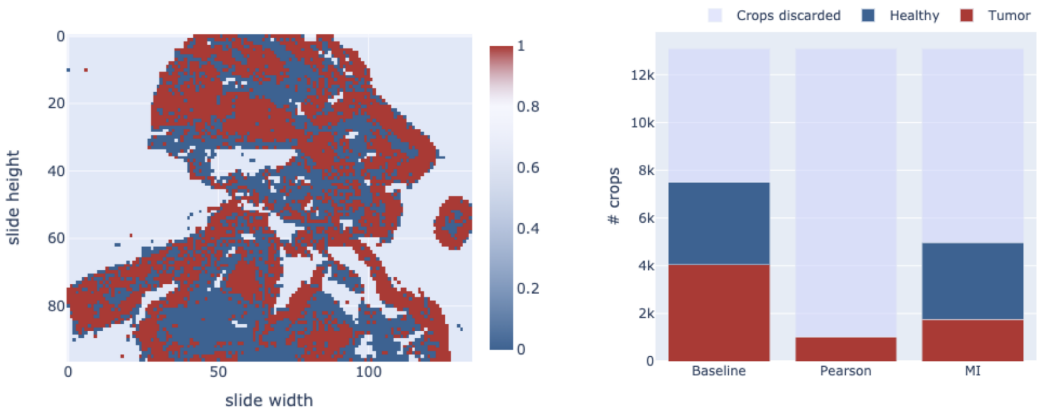

Figure A.18. On the left, the heatmap shows how the Baseline model classifies each crop on the slide. On the right is a stacked bar chart showing how the predictions vary in the Baseline, Pearson, and MI case.

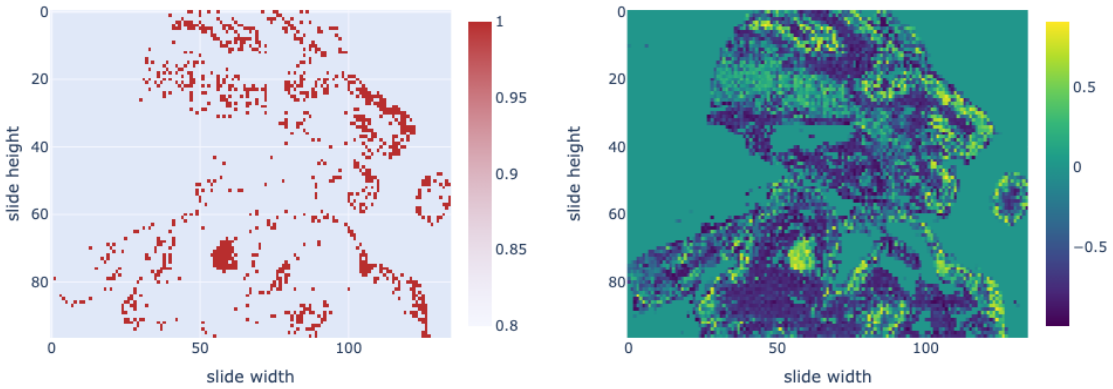

Figure A.19. Pearson case. On the left, the heatmap illustrates how the model classifies each crop on the slide, not considering sub-threshold correlation crops. On the right is shown the correlation value heatmap.

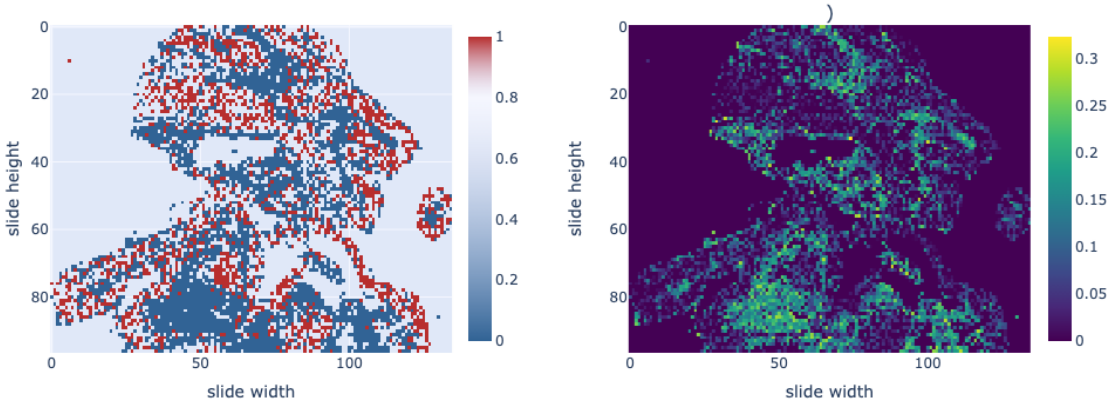

Figure A.20. MI case. On the left, the heatmap illustrates how the model classifies each crop on the slide, not considering sub-threshold correlation crops. On the right is shown the correlation value heatmap.

A.6 Healthy slide 3

The Pearson correlation succeeds in identifying areas of healthy tissue, but the classifier identifies those areas as tumor. What this example demonstrates is how correlation (Pearson's in this case) can be used to adjust classification.

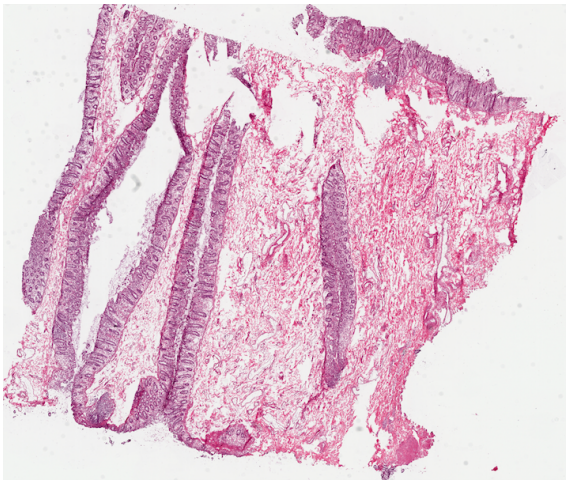

Figure A.21. Original image.

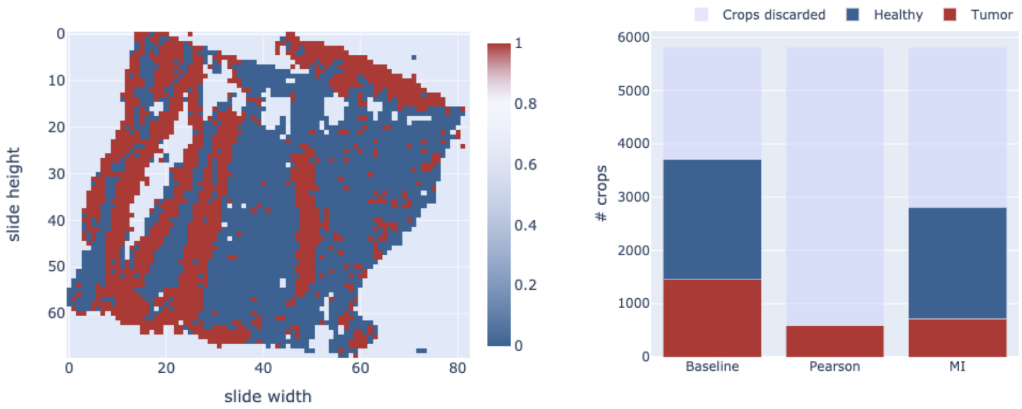

Figure A.22. On the left, the heatmap shows how the Baseline model classifies each crop on the slide. On the right is a stacked bar chart showing how the predictions vary in the Baseline, Pearson, and MI case.

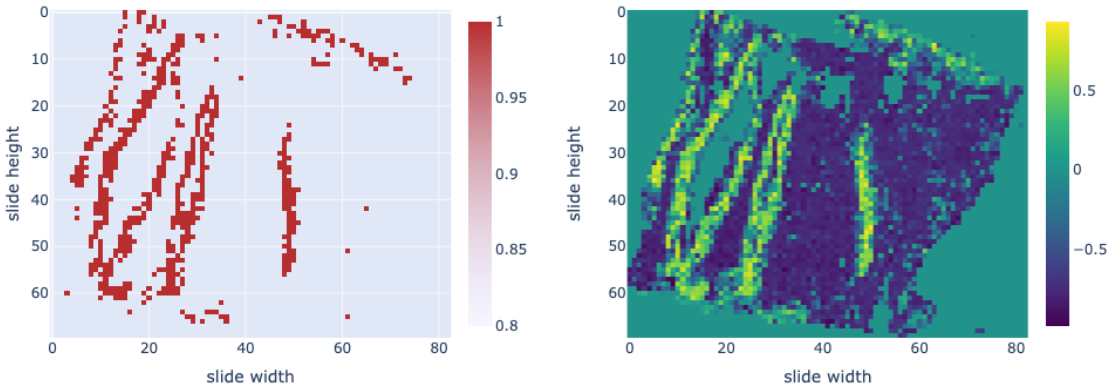

Figure A.23. Pearson case. On the left, the heatmap illustrates how the model classifies each crop on the slide, not considering sub-threshold correlation crops. On the right is shown the correlation value heatmap.

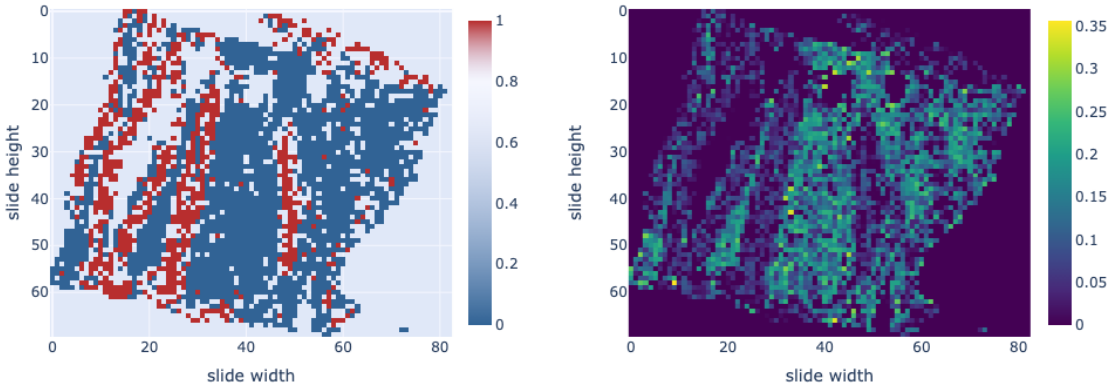

Figure A.24. MI case. On the left, the heatmap illustrates how the model classifies each crop on the slide, not considering sub-threshold correlation crops. On the right is shown the correlation value heatmap.

Bibliography

[1] Andreas Holzinger1, Benjamin Haibe-Kains, and Igor Jurisica. Why imaging data alone is not enough: Ai-based integration of imaging, omics, and clinical data. *European Journal of Nuclear Medicine and Molecular Imaging*, 46:2722–2730, 2019.

[2] Laura Antonelli, Mario Rosario Guarracino, Lucia Maddalena, and Mara Sangiovanni. Integrating imaging and omics data: A review. *Biomedical Signal Processing and Control*, 52:264–280, 2019.

[3] National Cancer Institute. Genomic data commons data portal. https://portal.gdc.cancer.gov/.

[4] University of Leeds. Virtual pathology project website. https://www.virtualpathology.leeds.ac.uk/.

[5] Francesco Ponzio, Enrico Macii, Elisa Ficarra, and Santa Di Cataldo. Colorectal cancer classification using deep convolutional networks - an experimental study. *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 58–66, 2018.

[6] Bray F., Ferlay J., Soerjomataram I., Siegel R.L., Torre L.A., and Jemal A. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68:394–424, 2018.

[7] Texas Digestive Desease Consultants. Colon cancer awareness. https://tddctx.com/colon-cancer-awareness/.

[8] National Cancer Institute. Colorectal cancer—patient version. https://www.cancer.gov/types/colorectal/patient/colon-treatment-pdq.

[9] Rebecca L. Siegel, Kimberly D. Miller, Ann Goding Sauer, Stacey A. Fedewa, Robert A. Smith, and Ahmedin Jemal. Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(3):145–164, 2020.

[10] Arnold M, Sierra MS, Laversanne M, and et al. Global patterns and trends in colorectal cancer incidence and mortality. *Gut*, 66:683–691, 2016.

[11] Manel Esteller. Epigenetics in cancer. *The New England Journal Of Medicine*, 358:1148–59, 2008.

[12] Mario F. Fraga, Esteban Ballestar, Maria F. Paz, Santiago Ropero, Fernando Setien, Maria L. Ballestar, Damia Heine-Suñer, Juan C. Cigudosa, Miguel Urioste, Javier Benitez, Manuel Boix-Chornet, Abel Sanchez-Aguilera, Charlotte Ling, Emma Carlsson, Pernille Poulsen, Allan Vaag, Zarko Stephan, Tim D. Spector, Yue-Zhong Wu, Christoph Plass, and Manel Esteller. Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, 102:10604–10609, 2005.

[13] Bob Weinhold. Epigenetics: The science of change. *Environmental Health Perspectives*, 114:160–167, 2006.

[14] Andrew P. Feinberg and Bert Vogelstein. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301:89–92, 1983.

[15] Marta Kulis and Manel Esteller. Dna methylation and cancer. *Advances in Genetics*, 70:27–56, 2010.

[16] Robin Holliday. The inheritance of epigenetic defects. *Science*, 238:163–170, 1987.

[17] Roberti A., Valdes A.F., Torrecillas R., and et al. Epigenetics in cancer therapy and nanomedicine. *Clinical Epigenetics*, 11, 2019.

[18] Susan E. Bates. Epigenetic therapies for cancer. *The New England Journal Of Medicine*, 383:650–663, 2020.

[19] Partha M. Das and Rakesh Singal. Dna methylation and cancer. *Journal Of Clinical Oncology*, 22:4632–4642, 2004.

[20] Keith D. Robertson and Alan P. Wolffe. Dna methylation in health and disease. *Nature Reviews Genetics*, 1:11–19, 2000.

[21] Sumei Wang and Wanyin Wu. Dna methylation alterations in human cancers. *Epigenetics in Human Disease (Second Edition)*, 6:109–139, 2018.

[22] Manel Esteller. Cancer epigenomics: Dna methylomes and histone-modification maps. *Nature Reviews Genetics*, 8:286–298, 2007.

[23] Melanie Ehrlich. Dna methylation in cancer: too much, but also too little. *Oncogene*, 21:5400–5413, 2002.

[24] Myoung Sook Kim, Juna Lee, and David Sidransky. Dna methylation markers in colorectal cancer. *Cancer and Metastasis Reviews*, 29:181–206, 2010.

[25] Stephen B Baylin. Dna methylation and gene silencing in cancer. *Nature Clinical Practice Oncology*, 2:S4–S11, 2005.

[26] Manel Esteller. Relevance of dna methylation in the management of cancer. *THE LANCET Oncology*, 4:351–358, 2003.

[27] Cindy D. Davis and Eric O. Uthus. Dna methylation, cancer susceptibility, and nutrient interactions. *Experimental Biology and Medicine*, 229(10):988–995, 2004.

[28] Michał W Luczak and Paweł P Jagodziński. The role of dna methylation in cancer development. *Folia Histochem Cytobiol*, 44(3):143–154, 2006.

[29] Jean-Pierre J. Issa. Dna methylation as a therapeutic target in cancer. *Clinical Cancer Research*, 13(6):1634–1637, 2007.

[30] Keith D Robertson. Dna methylation, methyltransferases, and cancer. *Oncogene*, 20:3139–3155, 2001.

[31] Oscar Jimenez del Toro, Sebastian Otálora, Mats Andersson, Kristian Eurén, Martin Hedlund, Mikael Rousson, Henning Müller, and Manfredo Atzori. Analysis of histopathology images from traditional machine learning to deep learning. *Biomedical Texture Analysis*, pages 281–314, 2017.

[32] Metin N. Gurcan, Senior Member, Laura E. Boucheron, Ali Can, Anant Madabhushi, Senior Member, Nasir M. Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE Reviews In Biomedical Engineering*, 2:147–171, 2009.

[33] Jonhan Ho, Anil V. Parwani, Drazen M. Jukic, Yukako Yagi, Leslie Anthony, and John R. Gilbertson. Use of whole slide imaging in surgical pathology quality assurance: design and pilot validation studies. *Human Pathology*, 37(3):322–331, 2006.

[34] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018.

[35] Stanley Cohen and Martha B. Furie. Artificial intelligence and pathobiology join forces. *The American Journal of Pathology*, 189(1):4–5, 2019.

[36] Mohamed Slaoui and Laurence Fiette. Histopathology procedures: From tissue sampling to histopathological evaluation. *Drug Safety Evaluation: Methods and Protocols, Methods in Molecular Biology*, 691:69–82, 2010.

[37] Brady Kieffer, Morteza Babaie, Shivam Kalra, and H.R.Tizhoosh. Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. *Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2017.

[38] Anant Madabhushi. Digital pathology image analysis: opportunities and challenges. *Imaging in Medicine*, 1:7–10, 2009.

[39] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[40] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. Pathol. Inf.*, pages 7–29, 2016.

[41] Dmitrii Bychkov, Nina Linder, Riku Turkki, Stig Nordling, Panu E. Kovanen, Clare Verrill, Margarita Walliander, Mikael Lundin, Caj Haglund, and Johan Lundin. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports*, 8, 2018.

[42] Zubair Ahmad1, Shabina Rahim1, Maha Zubair1, and Jamshid Abdul-Ghafar. Artificial intelligence (ai) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. a comprehensive review. *Diagnostic Pathology*, 16, 2021.

[43] Stephanie Robertson, Hossein Azizpour, Kevin Smith, and Johan Hartman. Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. *Translational Research*, 194:19–35, 2018.

[44] Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis C-A, and et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med*, 2019.

[45] F. d. A. Zampirolli, B. Stransky, A. C. Lorena, and F. L. d. M. Paulon. Segmentation and classification of histological images - application of graph analysis and machine learning methods. *23rd SIBGRAPI Conference on Graphics, Patterns and Images*, pages 331–338, 2010.

[46] Paola Sena, Rita Fioresi, Francesco Faglioni, Lorena Losi Giovanni Faglioni, and Luca Roncucci. Deep learning techniques for detecting preneoplastic and neoplastic lesions in human colorectal histological images. *Oncology Letters*, 18:6101–6107, 2019.

[47] Elene Firmeza Ohata, João Victor Souza das Chagas, Gabriel Maia Bezerra, Mohammad Mehedi Hassan, Victor Hugo Costa de Albuquerque, and Pedro Pedrosa Rebouças Filho. A novel transfer learning approach for the classification of histological images of colorectal cancer. *The Journal of Supercomputing*, 2021.

[48] Karren Dai Yang, Anastasiya Belyaeva, Saradha Venkatachalapathy, Karthik Damodaran, Abigail Katcoff, Adityanarayanan Radhakrishnan, G. V. Shivashankar, and Caroline Uhler. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nature Communications*, 12(31), 2021.

[49] Dongdong Sun, Ao Li, Bo Tang, and Minghui Wang. Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. *Computer Methods and Programs in Biomedicine*, 161:45–53, 2018.

[50] Nova F. Smedley and William Hsu. Using deep neural network for radiogenomic analysis. *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018.

[51] Xinliang Zhu, Jiawen Yao, Xin Luo, Guanghua Xiao, Yang Xie, Adi Gazdar, and Junzhou Huang. Lung cancer survival prediction from pathological images and genetic data - an integration study. *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1173–1176, 2016.

[52] Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, Neal I. Lindeman, and Faisal Mahmood. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 2020.

[53] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A. Gutman, Jill S. Barnholtz-Sloan, José E. Velázquez Vega, Daniel J. Brat, and Lee A. D. Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *National Academy of Sciences*, 115(13):2970–2979, 2018.

[54] Hao J., Kosaraju S.C., Tsaku N.Z., Song D.H., and Kang M. Pagenet: Interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. *Pac Symp Biocomput*, 25:355–366, 2020.

[55] Amal Katrib, William Hsu, Alex Bui, and Yi Xing. "radiotranscriptomics": A synergy of imaging and transcriptomics in clinical assessment. *Quantitative Biology*, 4(1):1–12, 2016.

[56] Chao Wang, Hai Su, Lin Yang, and Kun Huang. Integrative analysis for lung adenocarcinoma predicts morphological features associated with genetic variations. *Pacific Symposium on Biocomputing 2017*, 22:82–93, 2017.

[57] Ziming Zhang, Heng Huang, Dinggang Shen, and The Alzheimer's Disease Neuroimaging Initiative . Integrative analysis of multi-dimensional imaging genomics data for alzheimer's disease prediction. *Frontiers in Aging Neuroscience*, 6:260, 2014.

[58] N. F. Smedley and W. Hsu. Using deep neural networks for radiogenomic analysis. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1529–1533, 2018.

[59] Openslide python 1.1.2 documentation. https://openslide.org/api/python/.

[60] scikit learn. Machine learning in python. https://scikit-learn.org/stable/.

[61] Raffaele Martone. Machine learning based performance prediction of automotive microcontrollers. Master's thesis, Politecnico di Torino.

[62] BY G. E. HINTON and R. R. SALAKHUTDINOV. Reducing the dimensionality of data with neural networks. *Science*, 28:504–507, 2006.

[63] Sebastian Ruder. An overview of gradient descent optimization algorithms. 2017.

[64] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.

[65] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2015.

[66] Keras API reference. https://keras.io/api/.

[67] SciPy API. https://docs.scipy.org/doc/scipy/reference/.

[68] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review*, 69, 2004.

[69] Brian C. Ross. Mutual information between discrete and continuous data sets. *PLoS ONE 9(2)*, 2014.

[70] L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Probl. Peredachi Inf*, 23:9–16, 1987.

[71] Numata J., Ebenhöh O., and Knapp E.W. Measuring correlations in metabolomic networks with mutual information. *Genome Inform.*, 20:112–122, 2008.

[72] Thomas Benjamin Berrett. *Modern k-Nearest Neighbour Methods in Entropy Estimation, Independence Testing and Classification*. PhD thesis, University of Cambridge, 2017.

[73] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 8 edition, 2017.

[74] Wei Q and Dunbrack RL Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS ONE 8(7)*, 2013.