



## Multimodal representations of biomedical knowledge from limited training whole slide images and reports using deep learning

Niccolò Marini <sup>a,\*</sup>, Stefano Marchesin <sup>b,1</sup>, Marek Wodzinski <sup>a,c</sup>, Alessandro Caputo <sup>d,e</sup>, Damian Podareanu <sup>f</sup>, Bryan Cardenas Guevara <sup>f</sup>, Svetla Boytcheva <sup>g,h</sup>, Simona Vatrano <sup>e</sup>, Filippo Fraggetta <sup>e,i</sup>, Francesco Ciompi <sup>i</sup>, Gianmaria Silvello <sup>b</sup>, Henning Müller <sup>a,j</sup>, Manfredo Atzori <sup>a,k</sup>

<sup>a</sup> Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais), Sierre, Switzerland

<sup>b</sup> Department of Information Engineering, University of Padua, Padua, Italy

<sup>c</sup> Department of Measurement and Electronics, AGH University of Kraków, Krakow, Poland

<sup>d</sup> Department of Pathology, Ruggi University Hospital, Salerno, Italy

<sup>e</sup> Pathology Unit, Gravina Hospital Caltagirone ASP, Catania, Italy

<sup>f</sup> SURFSara, Amsterdam, The Netherlands

<sup>g</sup> Ontotext, Sofia, Bulgaria

<sup>h</sup> Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria

<sup>i</sup> Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands

<sup>j</sup> Medical faculty, University of Geneva, 1211 Geneva, Switzerland

<sup>k</sup> Department of Neurosciences, University of Padua, Padua, Italy

### ARTICLE INFO

#### MSC:

41A05

41A10

65D05

65D17

#### Keywords:

Computational pathology

Multimodal Learning

Medical ontology

Natural language processing

Colon cancer

### ABSTRACT

The increasing availability of biomedical data creates valuable resources for developing new deep learning algorithms to support experts, especially in domains where collecting large volumes of annotated data is not trivial. Biomedical data include several modalities containing complementary information, such as medical images and reports: images are often large and encode low-level information, while reports include a summarized high-level description of the findings identified within data and often only concerning a small part of the image. However, only a few methods allow to effectively link the visual content of images with the textual content of reports, preventing medical specialists from properly benefitting from the recent opportunities offered by deep learning models. This paper introduces a multimodal architecture creating a robust biomedical data representation encoding fine-grained text representations within image embeddings. The architecture aims to tackle data scarcity (combining supervised and self-supervised learning) and to create multimodal biomedical ontologies. The architecture is trained on over 6,000 colon whole slide Images (WSI), paired with the corresponding report, collected from two digital pathology workflows. The evaluation of the multimodal architecture involves three tasks: WSI classification (on data from pathology workflow and from public repositories), multimodal data retrieval, and linking between textual and visual concepts. Noticeably, the latter two tasks are available by architectural design without further training, showing that the multimodal architecture that can be adopted as a backbone to solve peculiar tasks. The multimodal data representation outperforms the unimodal one on the classification of colon WSIs and allows to halve the data needed to reach accurate performance, reducing the computational power required and thus the carbon footprint. The combination of images and reports exploiting self-supervised algorithms allows to mine databases without needing new annotations provided by experts, extracting new information. In particular, the multimodal visual ontology, linking semantic concepts to images, may pave the way to advancements in medicine and biomedical analysis domains, not limited to histopathology.

\* Corresponding author.

E-mail addresses: [niccolo.marini@hevs.ch](mailto:niccolo.marini@hevs.ch) (N. Marini), [stefano.marchesin@unipd.it](mailto:stefano.marchesin@unipd.it) (S. Marchesin).

<sup>1</sup> Both authors contributed equally to this work.

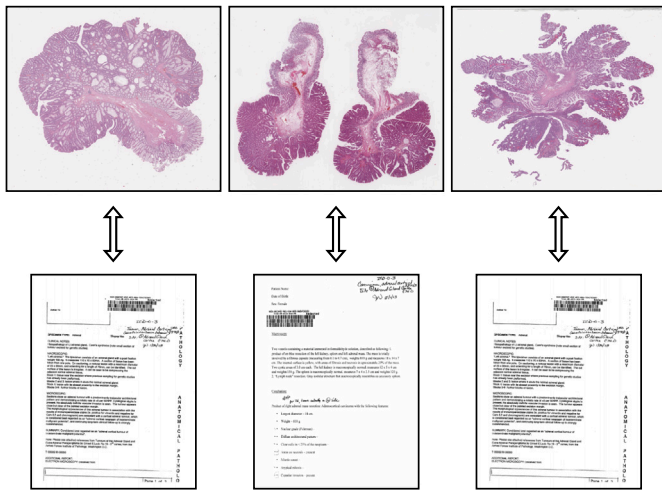


Fig. 1. Some examples of colon WSIs paired with the corresponding pathology reports.

## 1. Introduction

The increasing production of multimodal biomedical data empowers the development of new deep learning algorithms to analyze and represent data, especially in domains where data annotations are few and heterogeneity is high, such as the histopathology domain. Still, few methods allow extracting knowledge and linking information from different medical modalities in effective ways. This paper presents a multimodal architecture combining the visual information encoded within histopathology images with the semantics within textual reports.

The collection of different biomedical data modalities aims to gather relevant information on varying aspects linked to patient health to identify possibly dangerous conditions. Data analysis from multiple biomedical modalities requires the development of new deep learning algorithms integrating them. Multimodal learning (Bulten et al., 2022) involves the combination of information from multiple modalities, aiming to learn relationships between modalities to improve data representation (Stahlschmidt et al., 2022). Multimodal learning algorithms are becoming increasingly popular in machine learning for several reasons. First, the information included in different modalities is usually complementary since every medical modality generally provides information on specific aspects of the patient condition (Heiliger et al., 2022). Second, the cost of collecting multimodal biomedical data is becoming relatively low (Nagai et al., 2017; Gaziano et al., 2016), thanks to the increasing amount of initiatives pairing multimodal sources of information (Acosta et al., 2022). Third, combining multiple sources of information can bring benefits in a domain where collecting annotated datasets is time-consuming, as data are heterogeneous (Amal et al., 2022) and inherently multifaceted.

These characteristics are particularly relevant in the histopathology domain. Histopathology involves the analysis of tissue samples (Gurcan et al., 2009) to identify the microscopic findings characteristic of diseases such as cancer. Pathologists are the medical experts analyzing tissue sections, the exploration of which can be time-consuming. The analysis of histopathology samples does not usually rely on digital assistance in clinical practice, despite the growing digitization of tissue samples (Pallua et al., 2020; Fraggetta et al., 2017). Digital pathology involves digitizing and managing tissue specimens, called Whole Slide Images (WSI), acquired at high resolution and usually stored in a multi-scale format. WSIs are usually paired with pathology reports (Hanna et al., 2020), which include observations derived from manual WSI analysis for a patient, possibly using several WSIs. Fig. 1 shows examples of WSIs paired with reports.

The collection of large-scale image repositories and associated diagnoses are paving the way for the development of the computational

pathology domain (Marini et al., 2022; Abels et al., 2019), a domain involving the development of algorithms for the automatic analysis of WSIs. Even if computational pathology algorithms are becoming more and more accurate, some limitations still limit their adoption in clinical practice, such as the need for annotated data (Campanella et al., 2019); the lack of model generalization on unseen data due to data heterogeneity (e.g. in terms of tissue morphologies and color variations) (Tellez et al., 2019; Marini et al., 2023); the limited combination between WSIs and other medical modalities in network design.

In particular, combining multiple modalities, such as WSIs and reports, is still challenging because of the relationship between modalities. Analyzing a specific medical modality requires a specific architectural design (Acosta et al., 2022; Huang et al., 2020), therefore combining heterogeneous architectures may not be trivial. For instance, images are often analyzed with Convolutional Neural Networks (CNN) or Visual Transformers (ViT), while reports with pre-trained Large Language Models (LLMs). Furthermore, the relationship among modalities can influence how they combine: if a modality is subordinated to another, not all available multimodal learning frameworks can be adopted. For this reason, most applications that combine images and reports exploit reports to produce weak labels, which are then used as ground truth for the corresponding images (Marini et al., 2022; Marchesin et al., 2022).

This paper presents a multimodal architecture combining the low-level visual information encoded within WSIs with the high-level semantics stored within textual reports. The novelty of the paper involves both technical aspects and the possibility to create visual ontologies with biomedical data. Technical aspects involve the adoption of self-supervised algorithms (SSL) in a context where training data are scarce. Usually, SSL algorithms require large amounts of training data (Azizi et al., 2022; Campanella et al., 2023; Caron et al., 2021). However, the collection of a large amount of biomedical samples may not be trivial. For this reason, the multimodal architecture shows a peculiar design: it consists of two input branch encoders (which separately encode WSIs and pathology reports) and it is trained combining supervised with self-supervised learning. The encoders aim to generate an embedding vector for each input modality, that are aligned exploiting peculiar SSL loss functions and are processed by a shared projection head and a shared classifier. During training, the loss function includes both weak and SSL terms to (i) optimize the classification of images and reports and (ii) learn relationships among modalities. While classification requires supervised learning, the terms involved in relationship learning across modalities aim to build a strong multimodal histopathology representation space in a SSL fashion. This choice differs from classical SSL algorithms and Vision-Language models (VLM), where no annotations are required being based on the idea that large unannotated datasets can be collected. Therefore, the lack of large training dataset, usually hard to collect in biomedical domain, is compensated by adding weak supervision.

The multimodal architecture, based on SSL and weakly supervised learning, aims to create visual ontologies of biomedical data from limited training sets. Learning relationships between images and reports via SSL allows mining databases to discover new knowledge without the need for annotations by medical experts. The chosen concepts to match are fine-grained concepts, collected from the ExaMode Ontology (Menotti et al., 2023). The adoption of these concepts is not trivial: even if VLMs allow to link visual and textual concepts, most of the SOTA algorithms focus on broader concepts, such as the cancer type (Vorontsov et al., 2023; Lu et al., 2023a) or molecular subtypes (Filiot et al., 2023).

The generated multimodal histopathology data representation can serve as a strong backbone to address various other tasks, showing more robust performance than its unimodal counterparts. The analysis described in this paper targets over 6,000 colon WSIs and pathology reports collected from digital pathology workflows and over 1,000 WSIs collected from publicly available datasets. The architecture is evaluated

on three tasks: WSI classification (evaluated on pathology workflow and publicly available data), multimodal data retrieval (considering a modality as input to retrieve the other one), and linking visual and textual concepts. The latter two tasks are available by design, obtained without the need for further training, learned without any supervision.

Colon cancer was selected as a use case because it is the fourth most frequently diagnosed cancer globally and its diagnosis is complex, requiring the identification of multiple concepts, such as the presence of infiltrated glands, the presence of dysplasia and its grades, and the presence of polyps.

### 1.1. Related work

**Multiple Instance Learning.** Multiple Instance Learning (MIL) is currently the state-of-the-art framework to train weakly-supervised models (i.e., models trained using global labels) in the computational pathology domain (Campanella et al., 2019; Ilse et al., 2018; Lu et al., 2021; Hashimoto et al., 2020). MIL allows the organization of data as a bag of instances (Carbonneau et al., 2018), where data annotations include information about the whole bag but not on the single instances. MIL algorithms process the single instances and then aggregate them. In the computational pathology domain, a WSI represents a bag including patches (i.e., the instances), and the available annotations involve the entire WSI. Campanella et al. (2019) showed that applying MIL can lead to the development of models reaching almost perfect cancer vs. non-cancer predictions, showing the amount of data needed (around 10,000 per tissue use case) to reach AUC = 0.99. (Ilse et al., 2018) presented the Attention-Based Multiple Instance Learning (ABMIL), the first MIL framework embedding an attention network to learn weights are assigned to each patch based on its significance in the overall prediction. Javed et al. (2022) presented Additive-MIL (ADMIL), a MIL framework aiming to extend the MIL formulation to multiclass scenarios, embedding an attention-pooling layer with a channel for every output class. Lu et al. (2021) presented Clustering-constrained Attention Multiple Instance Learning (CLAM), an embedding-based MIL algorithm, adopting a cluster technique to aggregate relevant instances and identify relevant regions. Li et al. (2021a) presented Dual-Stream MIL, a MIL framework producing patch-level and image-level predictions. The instance-level predictions are evaluated only on a subset of relevant patches, aggregated using an attention mechanism to produce a WSI-level embedding, adopted to classify WSIs. Shao et al. (2021) presented TransMIL, a MIL framework combining CNN backbone and ViT components (Vaswani et al., 2017), to exploit spatial properties included within WSIs. The Transformer architecture represents instance features as a sequence of tokens, adopting a self-attention mechanism to highlight relationships between individual instances lost in attention networks. Li et al. (2021b) presented Deformable Transformer for Multiple Instance Learning (DTMIL), a hybrid architecture including convolutional layers and ViT components. The architecture allows focusing attention on a sub-set of relevant patches instead of the entire WSI, limiting the range of self-attention and requiring less computational power. Zhang et al. (2023) presented Multi-Level Multiple Instance Learning (MMIL-Transformer), a pure Transformer architecture to classify WSIs, that aims to mimic the behavior of pathologists combining the representations of small sub-regions of interest.

**Self-supervision.** Self-supervised learning (SSL) is a framework investigating how to exploit unlabeled data to learn a relevant data representation that can be fine-tuned afterward to perform specialized downstream tasks. The self-supervised domain is reaching increasing success, especially in domains where collecting annotated data is time-consuming (Koohbanani et al., 2021; Srinidhi et al., 2022), such as computational pathology. SSL algorithms aim to learn peculiar features and relationships from data collected from a specific domain instead of adopting natural images to pre-train the network. Currently, most of the algorithms adopted in computational pathology,

such as MoCo (He et al., 2020), simCLR (Chikontwe et al., 2020), and DINO (Caron et al., 2021), are contrastive algorithms adopted from general computer vision domains. The algorithms show similar characteristics since they aim to learn a data representation where similar samples are close to each other and far from dissimilar examples, aiming to minimize the distance (in the embedding space) between the embeddings representing similar samples and maximize the distance between the embeddings representing dissimilar samples. Couples of similar and dissimilar examples are automatically generated via data augmentation, so that a sample is similar to its augmented version (i.e., after a transformation such as rotation, flipping, or color perturbation) and dissimilar to other samples in a batch (He et al., 2020; Dehaene et al., 2020). Azizi et al. (2022) presented REMEDIS as a training strategy to improve the medical data representation. REMEDIS includes two training steps: a SSL pre-training strategy (on natural and then medical data), exploiting the simCLR algorithm and large datasets, and then a fine-tuning strategy, training the model on limited data. Chen et al. (2022) presented the Hierarchical Image Pyramid Transformer (HIPT), a ViT architecture trained on over 10,000 WSIs (from 33 cancer types) to capture tissue structures from multiple magnifications, combining two levels of self-supervised learning. Wang et al. (2022) presented Semantically-Relevant Contrastive Learning (SRCL) that aligns multiple positive instances with similar visual concepts instead of couples of positive examples collected from WSIs. The positive instances are generated using data augmentation and some semantically relevant images identified from a memory bank. Chen and Krishnan (2022) presented a study comparing different self-supervised and weakly-supervised strategies, aiming to identify the more robust representation in computational pathology, showing that ViT's pre-trained using the DINO algorithm (based knowledge distillation) guarantee robust and interpretable features. Filiot et al. (2023) presented iBOT, a self-supervised transformer-based algorithm based on Masked Image Modeling (MIM), pre-trained on over 40 million histopathology images. Campanella et al. (2023) presented a study including the largest histopathology dataset ever collected, over 3 billion images collected from over 423,000 WSIs. The study aims to compare the pre-training of ViT exploiting DINO and the masked autoencoder (MAE) algorithms, evaluating learned representation is evaluated on downstream tasks. Vorontsov et al. (2023) presented Virchow, a large neural network trained with a DINO algorithm on over 1.5 million WSIs.

**Vision-language models for image representation learning.** Vision-Language models (VLM) are algorithms aiming to build a data representation combining images and texts. VLMs are trained with images and the corresponding text (e.g., a textual description of the image content) to learn how to link the information from the two modalities. For this reason, VLM models are usually designed with multiple input branches including architectures to process specific input data, respectively, to input images and texts. The goal of VLMs is to build a stronger data representation that can be adopted as a backbone to solve downstream tasks, avoiding the need for annotations that may be expensive to collect. Radford et al. (2021) presented Contrastive Language-Image Pre-training (CLIP), a self-supervised algorithm to align visual and textual representations. CLIP is trained to maximize the similarity between corresponding image-text pairs while minimizing the similarity on unrelated pairs, exploiting a contrastive loss function (Oord et al., 2018), using around 400 million paired image-texts collected from publicly available repositories. Yu et al. (2022) presented Contrastive Captioner (CoCa), an algorithm to pre-train image-text encoder-decoder. CoCa is trained by combining a contrastive loss function with a captioning loss aiming to minimize the dissimilarity between the predicted sequence of output tokens and the actual target sequence. Zhang et al. (2020) presented Contrastive Visual Representation Learning from Text (ConVIRT), a framework aiming to learn image representations combining medical images and the corresponding reports. The training

aims maximize the similarity between image-to-text pairs and the similarity between text-to-image pairs. In both cases, the loss function is the Noise-Contrastive Estimation loss function (InfoNCE), which is asymmetric for each input modality. Wang et al. (2021) presented a UniFied TransFormer (UFO), a Transformer-based architecture that can be trained with both unimodal and multimodal data, depending on the task to solve. UFO includes a single encoder adopted for multiple modalities, that are concatenated. The encoder is trained to optimize several loss functions: an image-text contrastive loss, an image-text matching loss and a masked language modeling loss.

**Multimodal learning in computational pathology.** The application of multimodal learning algorithms in computational pathology aims to combine multiple sources of pathology data (usually histopathology images with reports or genomics data). Multimodal algorithms are usually pre-trained on agnostic tasks to provide a robust (including complementary characteristics) backbone model that can be further applied to specific downstream tasks. Lu et al. (2023b) presented MI-Zero, a framework to pre-train a model in a self-supervised fashion, exploiting a mechanism similar to CLIP. MI-Zero is pre-trained on many image-text pairs (around 33,000 histopathology image-report pairs) and then adopted to solve several downstream tasks, exploiting zero-shot learning. Furthermore, the text encoder is pre-trained using a corpora including around 900,000 reports from two hospitals and PubMed repositories. Huang et al. (2023) presented Pathology Language Image Pre-Training (PLIP) a vision-language model trained with a self-supervised algorithm to align histopathology images and the corresponding description. PLIP is pre-trained to adopt OpenPATH, a dataset including over 200,000 histopathology images paired with the corresponding text. Exploiting the alignment among modalities, PLIP can be adopted to classify samples labeled with unseen classes using zero-shot learning. Lu et al. (2023a) presented CONtrastive learning from Captions for Histopathology (CONCH), a visual-language foundation model designed to exploit several histopathology images and biomedical text. CONCH is trained using over 1.17 million image-text pairs, using two loss functions: an image-text contrastive loss to align the multimodal representation and a captioning loss function, such as the one proposed in CoCa.

## 1.2. Main contributions

In this work, we aim to address the following research question:

**RQ:** Can high-level concepts from textual reports be effectively combined with low-level image representations?

To do so, we present a multimodal architecture combining visual information from images with textual information from reports to improve histopathology data representations. Specifically, the contributions of this work are:

- A multimodal learning architecture combining images and reports, leading to a stronger histopathology data representation, that can be used as a backbone to solve computational pathology tasks.
- A multimodal histopathology data representation allows outperforming unimodal representations in terms of WSI classification, making it possible to exploit smaller datasets to train effective networks.
- The combination of self-supervised learning methods to learn similarities and dissimilarities between images and reports, using limited training datasets.
- The representation of medical ontologies in terms of visual knowledge, linking visual information from images with textual information from reports.

The rest of the paper is organized as follows: Section 2 describes the multimodal architecture, the dataset, including WSIs paired with reports, and the experimental setup. Section 3 describes the experimental results of the multimodal architecture on four tasks: WSI classification on pathology workflow data, Section 2.4.1, WSI classification on publicly available data, Section 2.4.1, multimodal data retrieval, Section 2.4.1, and the linking between visual and textual concepts, Section 2.4.1. Section 5 provides a discussion on the obtained results, while Section 6 concludes the paper with some final remarks.

## 2. Methods

### 2.1. Data

The dataset used in this paper includes over 6,000 colon WSIs and reports collected from the pathology workflows of two hospitals (the Catania cohort and Radboudumc) and over 1,000 colon images collected from two publicly available repositories.

Data from pathology reports are used to train and test the architecture. The architecture is trained with both WSIs and reports, while during the test phase, the model is mutually exclusive: only WSIs or reports can be used. Both images and reports are manually annotated by experts with five classes: Adenocarcinoma, High-Grade Dysplasia (HGD), Low-Grade Dysplasia (LGD), Hyperplastic Polyp and Normal Glands. The classes are not mutually exclusive, leading to multilabel annotations. Pathology workflow data are not manually selected, leading to unbalanced data in terms of class distribution, from both the Catania cohort and Radboudumc data. This choice aims to simulate a common scenario in digital pathology, where information about the image content is easy to be collected. In fact, querying a LIS for specific information about WSIs is often not feasible.

WSIs and reports are paired together. WSIs include gigapixel tissue samples, heterogeneous in terms of stain and sample types. Stain variability is a consequence of the heterogeneous acquisition procedures, especially regarding whole slide scanners and the composition of chemical reagents applied to the tissue. Catania cohort images are scanned using two Aperio scanners and two 3DHitech ones (at magnification 20/40x), while Radboudumc images are scanned using 3DHitech (at magnification 40x). Furthermore, the images are acquired with different types of medical tests: the Catania cohort dataset mainly includes colorectal polypectomies, biopsies, tissue resections and margin resections; while the Radboudumc dataset mainly includes biopsies and few polypectomies. Usually, biopsies are smaller in terms of size than colorectal polypectomies and tissue resections. The latter two types of images include more tissue and therefore more patches can be extracted. Table 1 includes a detailed composition of data collected from pathology reports, split in training and testing partitions.

Reports include heterogeneous free-text short descriptions of the findings identified during the analysis, that may lead to one or more high-level concepts linked to the considered classes (e.g., high-grade and low-grade dysplasia). The finding description is collected from the 'Conclusion' field, that may include macroscopic or microscopic description about the single WSI to which it is paired, avoiding any other information about the patient (such as the family history or personal information). Textual reports are heterogeneous in terms of source language, internal structure organization, and textual content. Reports are translated to English before the analysis, but the source languages are Italian and Dutch for the Catania cohort and Radboudumc, respectively. The internal structure of the reports varies according to the workflows from which they originate: The Catania cohort reports include a field with the diagnosis of the images, while Radboudumc reports include a field with the diagnosis of the whole block including one or more images. The last source of heterogeneity involves the textual content, as samples and reports are collected over the years. Therefore, several pathologists wrote the reports with their own personal style. Furthermore, Catania cohort reports are manually typed,

**Table 1**

Composition of the dataset collected from the pathology workflows of the Catania cohort and Radboudumc. The dataset includes paired WSIs and reports. The dataset is split into training and testing partitions. A 10-fold cross validation approach is applied to train and validate the models.

Source	Adenocarcinoma	HGD	LGD	Hyperplastic Polyp	Normal glands	Total
Training data						
Catania	893	774	1263	470	579	3091
Radboudumc	395	357	856	952	939	3085
Total	1288	1131	2119	1422	1518	6176
Testing data						
Catania	111	96	113	32	98	348
Radboudumc	75	65	146	119	193	520
Total	186	161	259	151	291	868

**Table 2**

Composition of the dataset collected from public available repositories: UNITOPatho and IMP-CRC. The dataset includes only images and it is split into training and testing partitions. A 10-fold cross validation approach is applied to train and validate the models.

UNITOPatho images						
Partition	HGD	LGD	Hyperplastic Polyp	Normal glands	Total	
Training	35	144	31	16	226	
Testing	11	40	10	5	66	
IMP-CRC images						
Partition	High-Grade Lesions		Low-Grade Lesions	Non-Neoplastic Lesions	Total	
Training	200		427	174	801	
Testing	51		100	52	203	

while Radboudumc ones are sometimes generated through “speech to text” tools, thus introducing an additional source of noise in the report analysis.

For what concerns WSIs from publicly available datasets, they are collected from two publicly available repositories: UNITOPatho (Barbano et al., 2021) and IMP-CRC (Oliveira et al., 2021). These repositories include heterogeneous images, in terms of color variations, sample types, and classes. UNITOPatho images are scanned using a Hamamatsu Nanozoomer S210 (at magnification 20x). They include sections of WSIs that are paired together and treated as a single WSI, and are annotated with four classes: High-Grade Dysplasia, Low-Grade Dysplasia, Hyperplastic Polyp, Normal Glands. IMP-CRC images are scanned using two Leica GT450 scanners (magnification 40x). They include colorectal biopsy and polypectomy slides, and are annotated with three classes: High-Grade Lesions, Low-Grade Lesions, Non-Neoplastic Lesions.

Publicly available data are collected to evaluate the capability of the multimodal data representation to generalize on heterogeneous data. The architecture is adopted to solve different classification problems (i.e. adopting other output classes and re-training the architecture classifier), comparing it with the unimodal data representation. Table 2 includes a detailed composition of data collected from publicly available repositories, split in training and testing partitions.

Overall, the data involved in this study presents a high degree of heterogeneity, which well resembles the landscape that these methods are required to face in real-case scenarios. In support of this, Fig. 2 shows the dataset heterogeneity, in terms of tissue sample type and of color distributions.

## 2.2. Multimodal architecture and training

We present a multimodal architecture to combine visual and textual information from biomedical images and textual reports, aiming to combine the low-level information from images with the high-level information from reports.

Fig. 3 provides an overview of the architecture. The architecture includes two input branches and some shared layers, including the classifier. The input branches process and encode WSIs and pathology reports. The image input branch (encoding WSIs) consists of a CNN

backbone and exploits the ADMIL framework to create an embedding of fixed dimension (128) representing a WSI. The embeddings are aggregated using an attention pooling layer, producing a single attention channel for each one of the classes involved in the classification (i.e. five). The attention channels are aggregated in a single embedding vector, representing the WSI, exploiting another attention pooling layer. The latter aggregation aims to create a single-dimension vector, that can be aligned to the one produced by the text encoder. The architecture is shown in the lower layer of Fig. 3. The text input consists of a BERT backbone, that outputs an embedding (768 in size) followed by a fully connected layer, to project the BERT embedding to a lower dimension size. The fully-connected layer creates an embedding of fixed dimension (i.e. 128, the same as WSI embeddings) representing a report. The embeddings of both modalities feed a shared projection layer. The choice of adopting a shared projection layer aims to tackle the lack of large datasets to train the architecture. Usually, VLMs embed a l2-normalized projection head for every modality, such as in CLIP architecture (Radford et al., 2021). However, those models are usually trained with a larger magnitude of samples (between hundreds of thousands and millions of samples) than our setup, where we adopted around 6,000 WSIs. The rationale behind this single projection head is to enhance alignment between image and text representations, given the limited data in self-supervised scenarios. This design aims to avoid overfitting, allowing the projection head to more effectively align modalities when processing both images and texts, leveraging its weight updates for classifying both modalities. The embeddings of both modalities feed the classifier, which outputs predictions on multilabel classes for both WSIs and reports. The embeddings representing WSIs and reports are not l2-normalized, as usually shown in VLMs, since the embeddings feed a classifier and it is not typical to normalize embeddings before a classifier.

The training of the network aims to classify histopathology samples (both WSIs and reports) and to combine high-level properties from reports with low-level properties from the image representation, based on raw pixels. While the first goal is achieved by minimizing the errors on the WSI and report predictions via two Binary Cross-Entropy loss function terms, the combination of high-level and low-level properties is achieved through several factors. First, the training loss function includes three terms to combine information: the NTXent loss function,

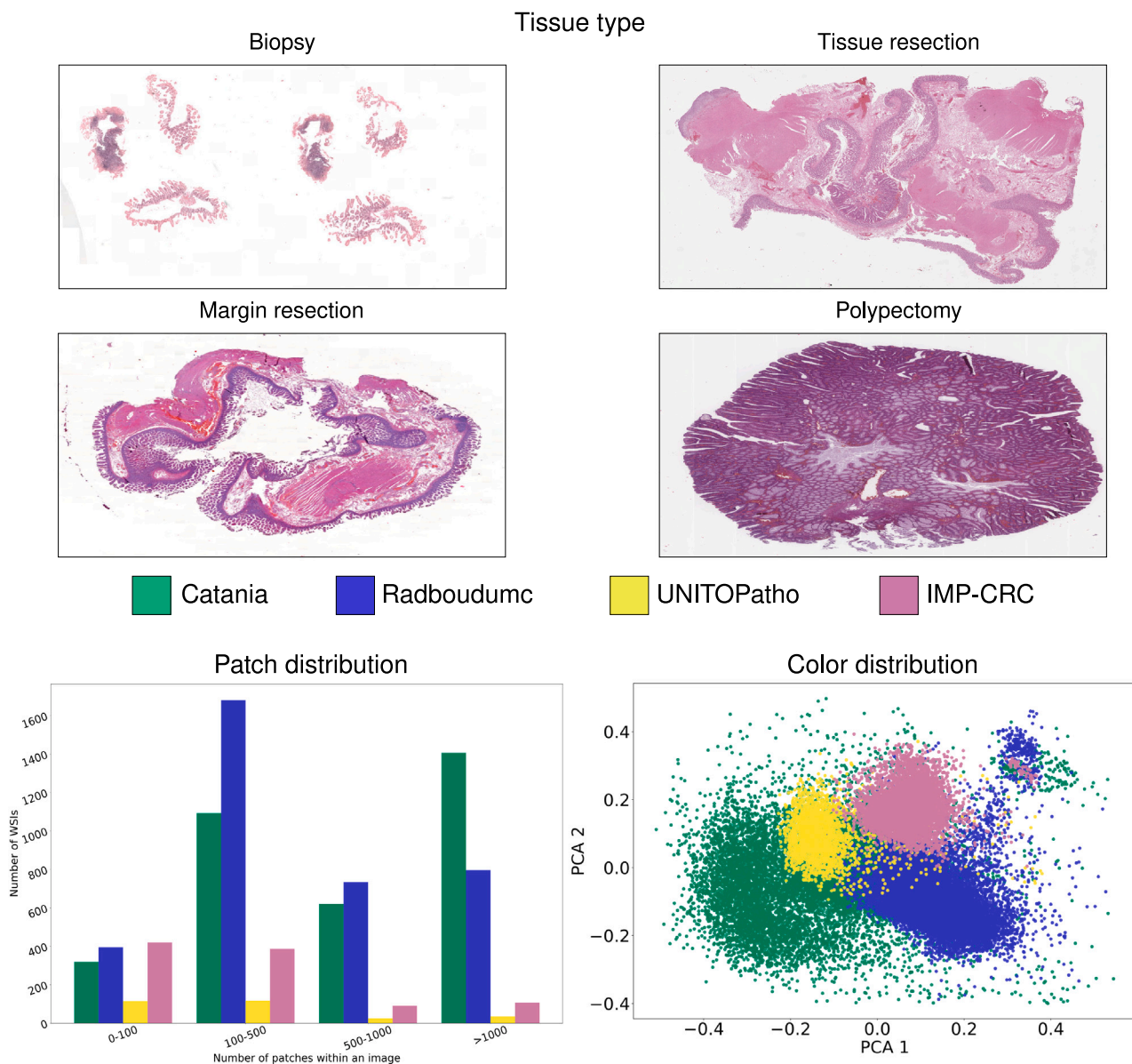


Fig. 2. Overview of dataset heterogeneity. The upper part includes examples of tissue samples: biopsy, tissue resection, margin resection and polypectomy. The lower part includes the distribution of patches per dataset (on the left), considering the Catania cohort (green), Radboudumc (Blue), UNITOPatho (yellow) and IMP-CRC; the distribution of color variation according the dataset (on the right), considering the PCA projection of the RGB components for H&E of the patches.

the L1-loss, and the cosine similarity loss. The ablation study involving the three loss functions is shown in Section 4. The NT-Xent loss function is a contrastive loss function adopted in self-supervised algorithms, such as MoCo and simCLR. During training, the architecture is fed with two input batches of dimension  $n$ : the WSIs and the reports. Every sample is linked to a similar example and  $n - 1$  dissimilar examples in the other batch. A similar example is the paired sample corresponding to the other modality (in the case of a WSI, the corresponding pathology report and viceversa), while instead the dissimilar examples are the other samples (i.e. in the case of a WSI, the reports of the other WSIs). The role of NT-Xent optimization is to learn the similarity and dissimilarity between samples in a batch. In this paper, the NT-Xent loss function is used to learn the similarity between a couple of corresponding WSI-report and the dissimilarity between unpaired WSIs and reports, as shown in a paper in recent papers about Vision-Language models, such as Radford et al. (2021), Lu et al. (2023b) and Huang

et al. (2023). NT-Xent is mainly sensitive to two parameters, which influence the loss function: a temperature parameter and the batch size. The temperature value chosen is fixed at 0.07, while the batch size is equal to 4. More information about the batch size is detailed in the Supplementary Material. The role of L1-loss and cosine similarity loss functions is to align the multimodal representations by minimizing the differences between the WSI and report representations. The L1-loss (Mean Absolute Error loss) is a function that minimizes the absolute differences between two vectors. The cosine similarity loss function is a function that minimizes the cosine of the angle between two vectors, computed as the dot product of the two vectors divided by the product of their magnitudes. The combination of both loss functions to align representations aims to avoid overfitting on training data (as shown in Section 4), since the direct minimization of the distance between multimodal representations, in combination with the relatively small training dataset (around 6,000 couples images-reports), could lead the

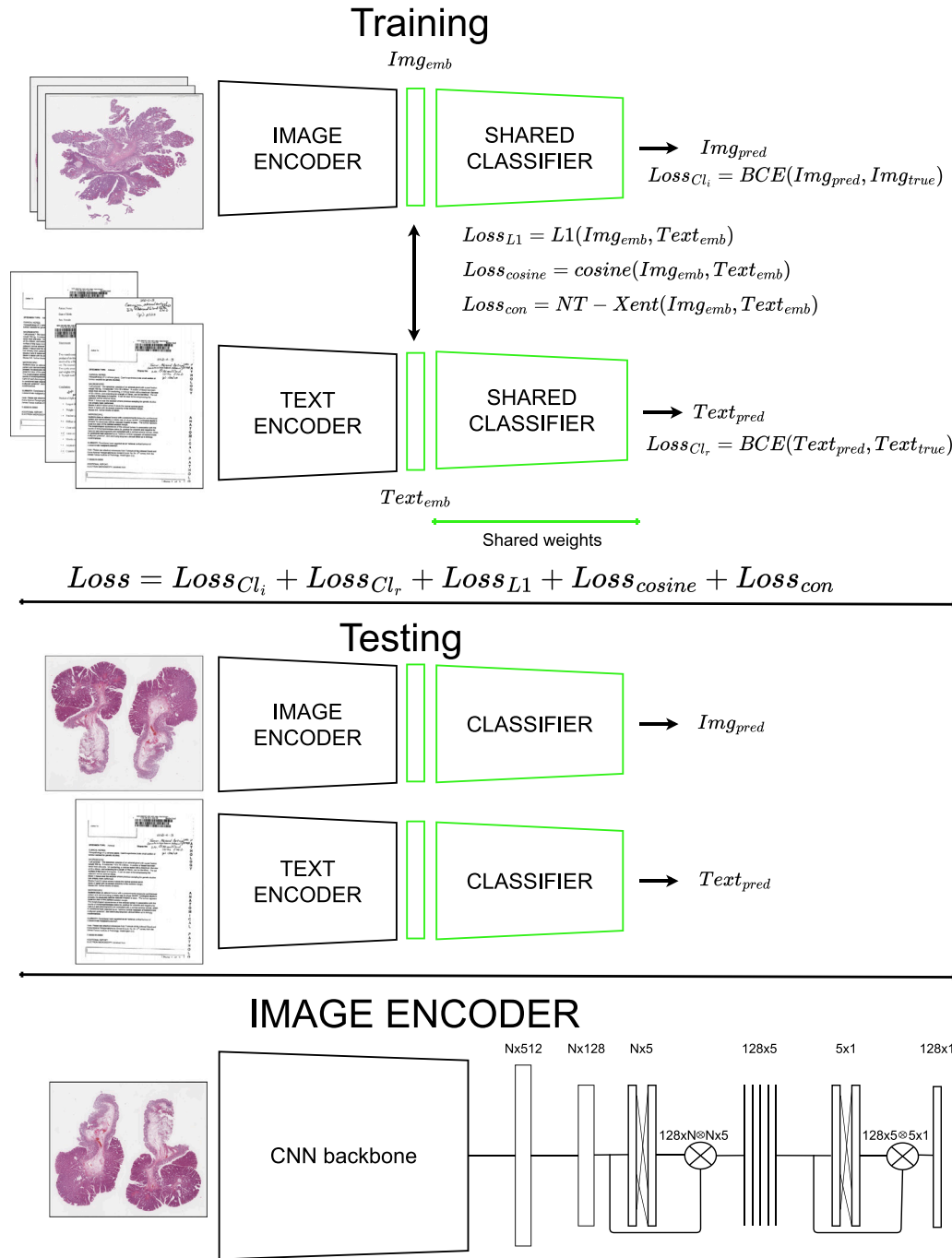


Fig. 3. Overview of the multimodal architecture. It includes two input branches, to encode WSIs and pathology reports, a shared projection head and a shared classifier. During the training both modalities are used, while instead during the testing the modalities are analyzed alone. The training involves the optimization of a loss function including several terms: the classification errors for both WSIs and reports, a self-supervised loss including NT-Xent loss, a L1-loss and a cosine similarity loss function. The self-supervised loss function aims to align the representations of WSI and reports. The lower part of the figure shows the image encoder: an ADMIL network, which includes a CNN backbone and two attention networks. The first attention network aggregates the single patches to create a WSI-embedding, containing an embedding for every class. In order to obtain a monodimensional embedding vector, another attention network aggregates the channels.

image and text representations into a small cluster in the embedding space (Liao, 2021). Furthermore, the fact that both modalities are classified by the same output branch helps to align the modality representations. In fact, during the training phase, the network weight updates influence both modalities, which are supposed to be as similar as possible.

In this paper, the multimodal architecture can be trained with several setups:

- *Unimodal*: the architecture is trained only on the classification of WSIs.
- *CLIP loss only*: the architecture is trained only on the self-supervised task, with the CLIP loss function.
- *Our SSL loss*: the architecture is trained only on the self-supervised task, with the loss function proposed in the paper (L1-loss + Cosine similarity Loss function + NT-Xent Loss).
- *Classification + CLIP loss*: the architecture is trained combining the classification terms and the CLIP loss function.

- *Classification + Our SSL loss*: the architecture is trained combining the classification terms and with the loss function proposed in the paper (L1-loss + Cosine similarity Loss function + NT-Xent Loss).

### 2.3. ExaMode ontology

The ExaMode ontology (Menotti et al., 2023) contains several components, including the key high-level concepts and properties for analyzing the findings identified during the tissue analysis.

The high-level concepts are summarized in three macro categories:

- Type of the polyp identified ('Adenoma-Serrated Polyps' or 'Malignant Polyps'). This kind of information refers to the node 'Polyp of Colon' and to its subclasses.
- Presence of Dysplasia, with the corresponding grade (low grade, medium grade, or high grade). This kind of information refers to the node 'Dysplasia' and to its subclasses.
- Characteristics of malignant polyps (considered as cancer), such as the type of tumor, the tumor grade. This kind of information refers to the node 'Adenocarcinoma' and to its subclasses.

### 2.4. Experimental setup

#### 2.4.1. Evaluation tasks

*WSI classification on pathology workflow data.* The multimodal architecture is tested on the classification of WSIs from pathology workflows, considering the unimodal and the multimodal setups (both with CLIP and our loss function to align multimodal representations). The testing partition includes an independent (split at patient-level) set of 868 WSIs, from both the Catania cohort (348) and Radboudumc (520). The goal of this task is to evaluate if the multimodal representation (learnt combining WSIs and reports) reaches higher performance compared with the unimodal representation (learnt only from images), using the same architecture. Furthermore, both architectures are trained with an increasing percentage of data (from 10% to the whole dataset) to evaluate if the combination of images and reports (paired together in pathology workflow) leads to high classification performance with a smaller amount of training data, reducing the need for the collection of large datasets.

*WSI classification on publicly available data.* The multimodal architecture is tested on the classification of images from public available datasets (UNITOPatho and IMP-CRC), considering the unimodal and the multimodal (our loss function) training setups. The goal of this task is to evaluate if the multimodal histopathology data representation can generalize better than unimodal one on heterogeneous data. The multimodal architecture is adopted as a pre-trained backbone to classify WSI from external datasets, including different classes. Therefore, WSI classification on publicly available data involves the finetuning of the multimodal architecture, trained using both couples of WSIs-reports or only WSIs, as shown in Paragraph 2.4.1. In this task, only the classifier is trained, while the backbone of the multimodal architecture image input branch is frozen. The classification performance of the multimodal representation is compared with the performance of the unimodal representation.

*Multimodal data retrieval.* The multimodal architecture is tested on the multimodal retrieval of images and reports, considering all the training setups, except the unimodal one. The task involves the retrieval of samples across modalities. When the input sample is an image, the goal is to retrieve the most similar reports; on the other hand, when the input sample is a report, the goal is to retrieve the most similar images. A point to stress about this task is that the multimodal retrieval task is inherently available after the multimodal training of the architecture without the need for additional training or finetuning, since the network is not directly trained on the retrieval task.

*Linking between visual and textual concepts.* The multimodal architecture is tested on the linking between visual and textual concepts, considering all the training setups, except the unimodal one. The concepts are the ones described in Section 2.3. The linking involves semantic concepts from pathology reports and the corresponding visual representations based on pixels, exploiting a zero-shot learning setup. Textual concepts can be extracted from textual reports, but it is still not completely clear how to link them to images, that include only raw-pixels, without any semantics. The linking involves the evaluation of similarity (cosine similarity) between the embeddings of images and concepts. In this paper, the images may be single patches ( $224 \times 224$  from magnification 10x) or entire WSIs. When the linking involves the single patches, the corresponding textual representation of every concept is evaluated, using the textual branch of the architecture. Concept embeddings are compared with the input patches via cosine similarity, linking the couples with the highest similarity value. One hundred patches per concept were selected (the ones with the highest similarity value), which an expert pathologist reviewed. When the linking involves the WSIs, the evaluation is slightly different since the task is not a multiclass problem, such as at patch-level, but it is rather a multilabel problem, since many concepts may be linked to a single WSI. Following a similar approach to the one proposed for patch linking, the cosine similarity between the visual embedding representing WSIs and the textual embedding representing concepts/classes is evaluated. Several thresholds are applied to link visual and textual concepts (0.7, 0.8, 0.9), as shown by Veeranna et al. (2016). When the cosine similarity exceeds the threshold, the textual concept is linked to the WSI.

The concept matching task is inherently available after the multimodal training of the architecture without the need for additional training costs or fine-tuning, since the network is not directly trained on the linking between visual and textual concepts task.

#### 2.4.2. Image pre-processing

Image pre-processing involves splitting the image into patches, selecting the ones from tissue regions, and discarding regions from the background. The splitting of WSIs into patches is necessary due to the gigapixel nature of WSIs. Currently, GPU hardware has limited memory and struggles to handle large input images. Images are split into patches of  $224 \times 224$  pixels, extracted from magnification 10x, using the Multi\_Scale\_Tools library (Marini et al., 2021b). The patch size is chosen considering that the ResNet34 backbone used as CNN requires this input data size. The magnification level is chosen considering that the WSIs at 10x allow visualizing the components that correctly identify the considered classes. Patches coming from background regions are not informative for analysis and are therefore discarded. The tissue and background regions are identified by generating tissue masks with HistoQC tool (Janowczyk et al., 2019).

#### 2.4.3. Report pre-processing

The report pre-processing involves the translation of the reports into English and the splitting of text into tokens. Reports are originally stored in Italian and Dutch, according to the workflows from which they originate. However, state-of-the-art NLP algorithms are often developed to use English. Therefore, the report content is first translated to English using pre-trained MarianMT neural machine translation models (Junczys-Downmunt et al., 2018). Then, before the analysis, translated reports are WordPiece tokenized using the BERT model vocabulary. BERT vocabulary includes around 30,000 tokens divided into words, subwords, or characters. When words are not included in the vocabulary, the WordPiece tokenizer divides words in known subword units or characters. By design, BERT accepts sequences of a maximum of 512 tokens, where the first token of each sequence is a special classification token ([CLS]) and the last is a special separator token ([SEP]). The final hidden state corresponding to the [CLS] token is the aggregate sequence representation for classification tasks (Devlin et al., 2018).



#### 2.4.4. Multimodal architecture pre-training

Both input branches for encoding images and reports are pre-trained.

The image input branch is a ResNet34 backbone, pre-trained with MoCo v2, a SSL framework aiming to learn features from input data. MoCo v2 is adopted to pre-train the CNN backbone on learning similarities and dissimilarities between input samples. In this paper, the input samples correspond to the WSI patches collected from the training partition. The concepts of similarity and dissimilarity among input data are achieved using data augmentation: an input sample is considered similar to its augmented version and dissimilar from the others. The augmented samples are collected in a queue to produce dissimilar examples from the input data. This paper's queue includes 16,384 samples, while each batch includes 256 samples. The augmentation pipeline includes several operations, implemented with Albumentations python library (Buslaev et al., 2020): horizontal and vertical flipping, random rotations (90/180/270 degrees), HUE saturation value, RGBShift, Contrast Limited AHE (CLAHE), random brightness, random contrast, elastic transformation, grid distortions. Each operation is applied to an input sample with a probability of 0.8. During CNN pre-training, a H&E-adversarial optimization (Marini et al., 2021a) is adopted to force the network to learn stain-invariant features to improve the capability of the network to generalize well when tested on data, including different stains that the ones included in the training partition.

The report input branch uses PubMedBERT (Gu et al., 2021) as the backbone, a BERT model pre-trained from scratch using abstracts from PubMed and full-text articles from PubMed Central (Gu et al., 2021). PubMedBERT has been pre-trained from scratch, using a domain-specific vocabulary, over biomedicine text to overcome the limitations of mixed-domain pre-training strategies. PubMedBERT has been optimized via Masked Language Modeling (MLM), Next Sentence Prediction (NSP), and adversarial pre-training, which introduces perturbations in the (input) embedding layer that maximizes the adversarial loss. Adversarial pre-training forces PubMedBERT to optimize the standard training objective (i.e., MLM and NSP) and minimize the adversarial loss (Liu et al., 2020). In this work, we take PubMedBERT as is and use it as an encoder for textual reports.

#### 2.4.5. Image data augmentation pipeline

Input whole slide images are augmented during CNN training using the Albumentations library (Buslaev et al., 2020). Image data augmentation pipeline includes three operations: 90/180/270 degrees random rotation, vertical and horizontal flipping, and hue-saturation-contrast (HUE) color augmentation. The augmentation is applied at image-level to have a consistent transformation for all the patches included in a WSI, with a probability of 0.5.

#### 2.4.6. Report data augmentation pipeline

The report data augmentation pipeline exploits three operations. The first operation is a back2back translation – that is, translating the input text to a different language and then back to English – using French, Italian, German, Spanish, Turkish, Chinese, Japanese, and Russian as middle languages, implemented using the nlpaug library (Ma, 2019). The second operation is an insert/rephrase strategy, implemented via the nlpaug library, which consists of slightly modifying the sentence by inserting new words or paraphrasing it. The third operation is a ChatGPT augmentation. The augmented report is produced by the GPT v3 model (text-davinci-003 backend) (Brown et al., 2020), submitting a prompt stating to modify the input text without changing its global content. The prompt adopted is: “Generate a different version of the following pathology report, without changing its content, but rather the order of the words, the sentences, and the medical terminology. [...]”

#### 2.4.7. Metrics to evaluate the architecture

The model's performance is evaluated on three tasks: the WSI classification (pathology workflow and publicly available data), multimodal data retrieval, and linking visual and textual concepts.

**WSI classification.** The model's performance on WSI classification is evaluated using the weighted macro F1-score. WSI classification is defined as a multilabel problem (on the pathology workflow data) and as multiclass problem (on the publicly available data). Weighted macro F1-score is adopted to tackle the class imbalance. The macro-weighted average involves the evaluation of the F1-scores for the single classes, which are then weighted according to the class support (number of true samples for the class). F1-score metric is defined as the average among the recall and the precision. The precision measures the capability of a classified to not misclassify negative samples as positive ones, while the recall measures the capability to classify positive samples correctly. The weighted F1-score is reported in terms of average and standard deviation of the ten experiment repetitions, evaluated on the test partition.

**Multimodal data retrieval.** The performance of the model on multimodal retrieval is evaluated using the precision@ $k$  and the mean average precision (mAP). The precision@ $k$  evaluates the number of relevant items retrieved concerning the total number of retrieved items (i.e.,  $k$  value). The paper's  $k$  value (i.e., cut-off) for the precision equals 5, 10, 50. The choice of adopting low cut-off values is based on the fact that medical experts querying the retrieval system may not have a large amount of time to check all possible outcomes, so the system must be effective with a small number of retrieved samples. The mean average precision involves the evaluation of the average precision (the area under the precision–recall curve) for the single input samples of the retrieval system. The average precisions are then averaged. The area under the recall-precision curve involves the evaluation of the recall@ $k$  and the precision@ $k$ , where  $k$  is the number of samples that can be retrieved (in this case the whole dataset from pathology workflows). In the paper, the  $k$  value for the mAP equals 1,000. The cut-off value for this metric is higher than the one adopted for precision@ $k$ . This choice is driven by the fact that the mAP metric also involves the evaluation of recall@ $k$ , which measures the ability to retrieve as many relevant documents as possible; therefore, a high cut-off value shows how well the system performs in terms of retrieving relevant content across a wide range of possibilities. Since data are multilabel, both metrics are computed with the micro-average (true positives and true negatives are considered separately and then aggregated).

**Linking between visual and textual concepts.** The performance of the model on the linking between visual and textual concepts is evaluated for each task using the accuracy (true positives over the total amount of samples).

### 2.5. Statistical significance test

The Wilcoxon Rank-Sum test (Woolson, 2007) is adopted during a comparison between two algorithms to verify if the performance difference is statistically significant ( $p$ -value < 0.05).

#### 2.5.1. $k$ -fold cross validation

The multimodal architecture is trained using  $k$ -fold cross-validation to show the robustness of the model concerning data used for training. The training partition is divided into  $k$  groups (in this paper  $k = 10$ ): for every training repetition,  $k-1$  groups are used to train the model, while the other group is used to validate the model. Data are divided at the patient level so that the samples collected from a patient cannot be on both training and validation partitions. During the training, the loss function involves optimizing five terms, as shown in Section 2.2. During the validation, the weights of the best model are stored when the loss function reaches the lowest value, considering only the WSI classification term of the loss function.

**Table 3**

Results for the performance of the multimodal architecture on the classification of WSIs, considering the Catania cohort and Radboudumc datasets. The evaluation involves the unimodal and the multimodal representation (considering both CLIP loss and our SSL loss function). The performance is evaluated with weighted F1-score, reporting the average and the standard deviation (of the models involved in the k-fold cross-validation) and including cumulative results for each dataset.

F1-score	Catania	Radboudumc	Cumulative
Unimodal	0.765 ± 0.013	0.771 ± 0.012	0.769 ± 0.011
Classification + CLIP loss	0.785 ± 0.013*	0.782 ± 0.013*	0.784 ± 0.013*
Classification + Our SSL loss	<b>0.788 ± 0.011*</b>	<b>0.792 ± 0.012*</b>	<b>0.790 ± 0.009*</b>

\* The results that are statistically significant (compared with the unimodal representation, using the Wilcoxon Test) are reported.

### 2.5.2. Hardware & software

The experiments are implemented using Python libraries. The deep learning algorithms are implemented and trained using PyTorch 1.11. Transformers 4.6.1 (Wolf et al., 2019) is used to implement the BERT architecture and to pre-process textual reports. Deep learning experiments are performed on a Tesla V100 GPU. WSIs are accessed using openslide 3.4.1 (Goode et al., 2013). The model performance is evaluated with the metrics implemented by scikit-learn 0.23. The image pre-processing and augmentation are applied using albumentations 1.8 (Buslaev et al., 2020).

### 2.5.3. Hyperparameters

The grid search algorithm is adopted to identify the optimal configuration of CNN hyperparameters. The optimal configuration chosen reaches the lowest loss function on WSI classification in the validation partition. The parameters involved in the grid search algorithm are: the optimizer (Adam selected; Adam and SGD tested); the number of epochs to train the model (15; above this amount of epochs, the validation partition loss function does not reach a lower level); the learning rate ( $10^{-3}$ ;  $10^{-2}$ ;  $10^{-3}$ ;  $10^{-4}$ ;  $10^{-5}$  were tested); the decay rate ( $10^{-3}$ ;  $10^{-2}$ ;  $10^{-3}$ ;  $10^{-4}$ ;  $10^{-5}$  were tested); the number of nodes in the intermediate layer after the ResNet backbone (128; 32, 64, 128, 256 were tested).

## 3. Results

**Multimodal representation latent space.** Fig. 4 shows the latent spaces including the embeddings representing either WSIs and reports, as the outcome of the unimodal representation (architecture trained only to classify WSIs, upper Figure) and the multimodal representation (architecture trained to classify samples and using our SSL loss function, lower Figure). Each dot in the Figure shows the embeddings, either WSIs (red) and reports (blue) evaluated on the internal testing partition (data from the Catania cohort and Radboudumc). The latent space is obtained via dimensionality reduction, pre-processing the embeddings with Principal Component Analysis (PCA) and then applying t-distributed stochastic neighbor embedding (t-SNE).

The upper part of Fig. 4 shows the embeddings representing WSIs and reports, outcome of the unimodal representation. Since the input data encode different characteristics and are processed with different architectures, the corresponding embeddings are separated in the space.

The lower part of Fig. 4 shows the embedding representing WSIs and reports, outcome of the multimodal representation. The multimodal representation is learned by forcing the representations of WSIs to be similar to the corresponding report representations and by learning relationships of similarity and dissimilarity among WSIs and reports, as shown in Section 2.2. The main characteristic of the multimodal representation is that WSI embeddings (red dots) are close to report embeddings (blue dots), showing that the image representation encodes both raw-pixel and concepts information. This characteristic is not trivial, since the characteristics of both modalities are complementary: while reports are short and include high-level concepts, images are large and the pixels do not include any semantic information.

**Table 4**

Results for the performance of the multimodal architecture on the classification of WSIs, considering the UNITOPatho and the IMP-CRC datasets. The evaluation involves the unimodal and the multimodal representation (considering both CLIP and our SSL loss function). The performance is evaluated with a weighted F1-score, reporting the average and the standard deviation (of the models involved in the k-fold cross-validation). The evaluation involves the unimodal and the multimodal histopathology data representation.

F1-score	UNITOPatho	IMP-CRC
Unimodal	0.790 ± 0.017	0.874 ± 0.018
Classification + CLIP loss	0.750 ± 0.062	0.881 ± 0.007
Classification + Our SSL loss	<b>0.824 ± 0.022*</b>	<b>0.894 ± 0.014*</b>

\* The statistically significant results (compared with the unimodal representation, using the Wilcoxon Test) are reported.

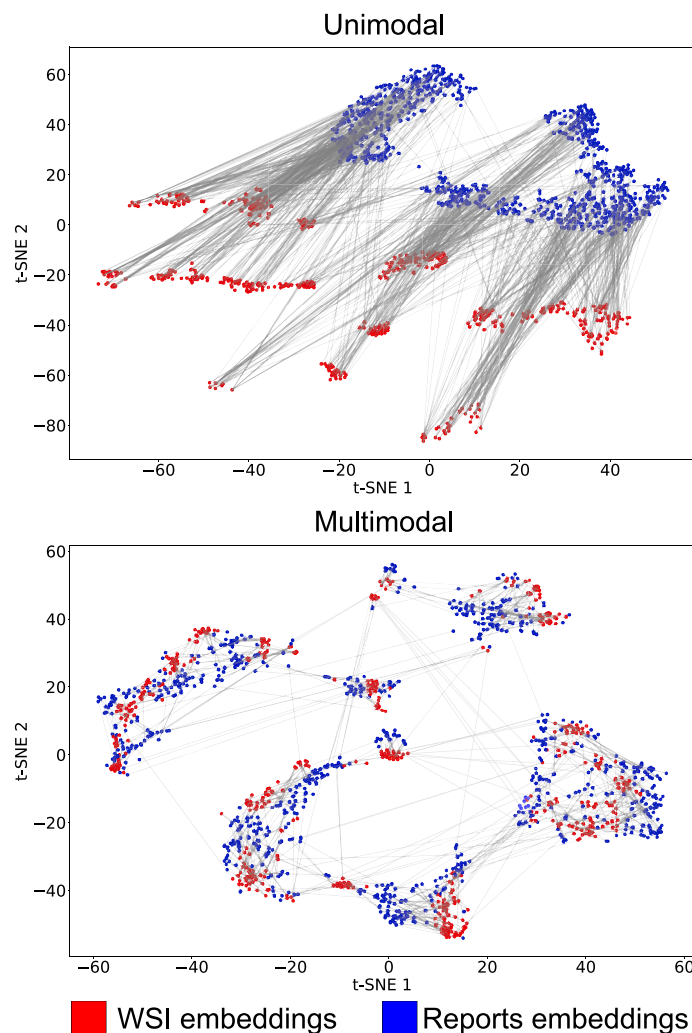
**WSI classification on pathology workflow data.** The multimodal representation outperforms the unimodal representation in the classification of WSIs collected from pathology workflows for every percentage of training data adopted in every testing partition.

The architecture is evaluated with the weighted F1-score, considering different percentages of input data and reporting the average and the standard deviation of ten models (trained using cross-validation). The architecture is trained with an increasing percentage of data, starting from 10% and up to the whole dataset. The training results on the increasing amount of data are shown in Fig. 5. The architecture trained with both modalities (using our loss function to align images and reports), but using half of the dataset (3,000 couples of WSIs-reports instead of 6,000), reaches the same performance of the unimodal architecture trained with the whole dataset. This suggests that combining images and reports may alleviate the need to collect large datasets to reach robust performance. When the whole training dataset is used (100%), the multimodal representation reaches higher accuracy than the unimodal one, with cumulative (considering both the Catania cohort and Radboudumc test partition) F1-score =  $0.790 \pm 0.009$ . In contrast, the unimodal representation reaches a cumulative F1-score =  $0.769 \pm 0.011$ , as shown in Table 3.

**WSI classification on publicly available data.** The multimodal representation, trained with our loss function to align the image and textual representations, outperforms the unimodal representation on classifying images from publicly available datasets. Both the representations are learned during the previous task (i.e., WSI classification on pathology workflow data, Section 2.4.1).

Images from publicly available datasets, collected from UNITOPatho and IMP-CRC, are annotated with different classes. While the architecture is originally trained with multilabel data (five classes), both UNITOPatho and IMP-CRC include multiclass data (respectively four and three classes). The architecture backbone, considering both the multimodal, trained with CLIP loss and our loss function to align modalities, and the unimodal representation, is pre-trained on pathology workflow data, as shown in Paragraph 2.4.1. On the other hand, the classifier is trained from scratch on different classes.

The multimodal representation reaches an F1-score =  $0.824 \pm 0.022$  on UNITOPatho dataset and an F1-score =  $0.894 \pm 0.014$  on IMP-CRC dataset, while the unimodal representation reaches respectively



**Fig. 4.** Overview of the latent space, considering both the unimodal representation (architecture trained only to classify WSIs, upper Figure) and the multimodal representation (architecture trained to classify samples and using our SSL loss function, lower Figure). The latent space of the multimodal representation shows regions where images and reports are close to each other. The latent space of the unimodal representation shows two separate clusters, one including images, the other including textual reports.

an F1-score =  $0.790 \pm 0.01$  on UNITOPatho dataset and an F1-score =  $0.874 \pm 0.018$  on IMP-CRC dataset. The multimodal representation trained with our SSL loss function also outperforms the multimodal representation obtained by aligning images and texts with CLIP loss. [Table 4](#) summarizes the results on publicly available datasets.

**Multimodal data retrieval.** The multimodal architecture allows building a robust retrieval system without any peculiar architectural design for reports and image retrieval compared with other methods developed for this purpose.

Due to the multimodal representation characteristics, the retrieval system can retrieve reports using a WSI input sample and WSIs using a report input sample. The system is evaluated using the precision@ $k$  (cut-off  $k$  chosen values are 5, 10, 50) and the mean average precision (mAP). In both cases, the performance is evaluated considering the samples from the testing partition of the Catania cohort and Radboudumc as input queries, with the possibility to retrieve data from all the WSIs and reports collected from pathology workflows (6176 samples from the Catania cohort and Radboudumc).

[Table 5](#) summarizes the performance reached on the multimodal retrieval, for both images and reports. The highest results are achieved by the combination of classification and SSL losses (CLIP and our SSL loss function to align modalities) that exceed both setups with only SSL losses. In particular, the multimodal setup combining classification and our SSL loss to align modalities reaches the highest performance in

terms of precision@ $k$ , for both datasets and modalities; while instead, the combination of classification and CLIP loss function reaches the highest performance in terms of mAP, for both datasets and modalities. These results highlight the importance of the classification term in the multimodal architecture training, especially when training datasets are not large in size.

**Linking between visual and textual concepts.** The multimodal representation allows linking tissue morphologies from images to high-level concepts included in the reports, building a multimodal knowledge graph of paired histopathology visual and text information.

The concepts adopted for the linking are extracted from the ExaMode colon ontology, presented in [Section 2.4.1](#). The images associated with the ontology concepts were reviewed by an expert pathologist. Visual concepts can be linked to textual concepts at patch-level and WSI-level.

[Table 6](#) shows the results of the multimodal architecture on the linking between visual and textual concepts, collected from reports, at patch-level.

[Table 6](#) shows the results of the linking between visual and textual concepts, collected from reports. The linking is effective, considering that the global accuracy of the multimodal architecture (trained combining classification and our SSL loss) is 0.607, higher than the other setups. The result is relevant, considering the fact that modality combination does not require any supervision and that most of the concepts

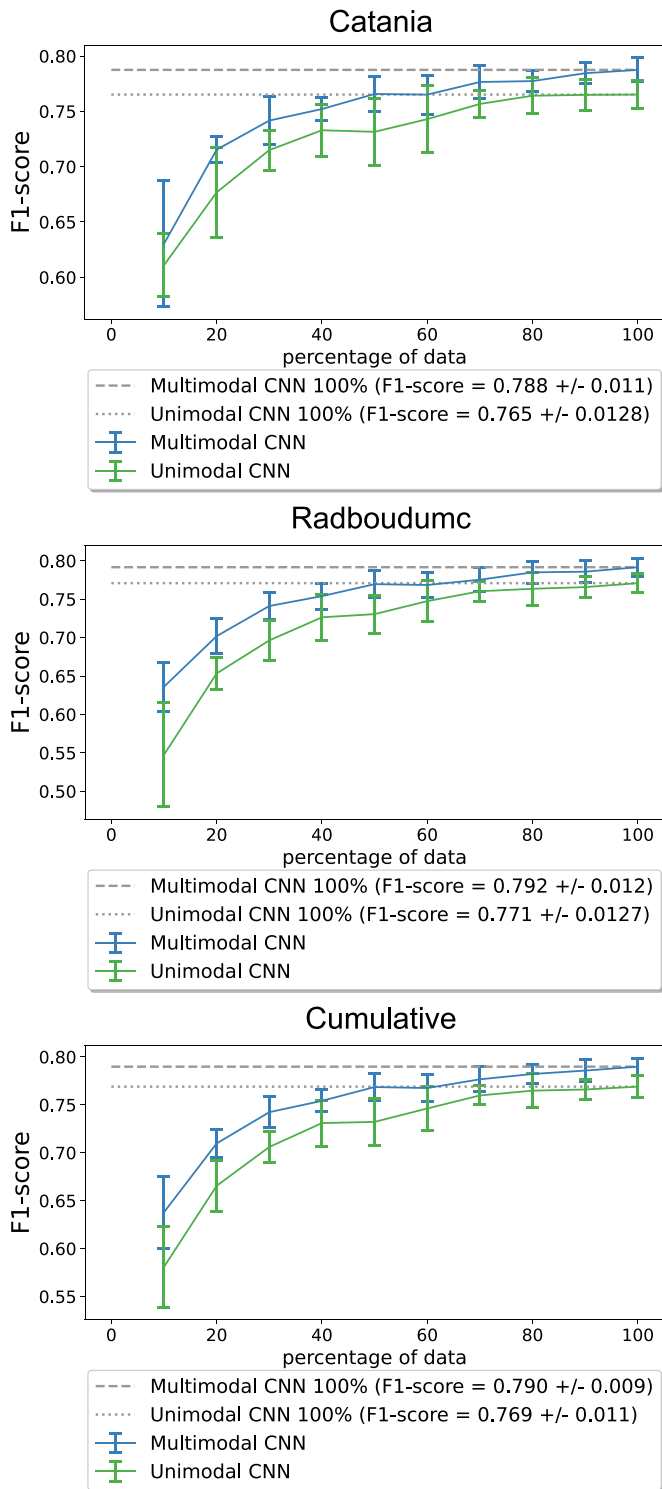


Fig. 5. Results for the average performance of the architecture. The evaluation involves the multimodal representation (blue line, classification + our SSL loss function) and the unimodal representation (green). The results are reported for the Catania cohort testing partition (upper sub-figure), the Radboudumc testing partition (middle sub-figure), and the combination of both testing partitions (bottom sub-figure). In the three sub-figures, the dashed line represents the performance of the architecture when trained by combining both images and reports on the whole dataset (100%); the dotted line represents the performance of the architecture when trained with only images on the whole dataset (100%).

Table 5

Overview of the results of the multimodal retrieval task. The task is evaluated on the image retrieval (reports as input) and the report retrieval (images as input), considering several multimodal methods: CLIP loss only, our SSL only, classification + CLIP loss, classification + our SSL loss.

CLIP loss only		
Retrieve reports (image as input)		
Metric	Catania	Radboudumc
Precision@5	0.255 ± 0.030	0.244 ± 0.097
Precision@10	0.253 ± 0.031	0.251 ± 0.076
Precision@50	0.249 ± 0.049	0.268 ± 0.064
mAP	0.179 ± 0.024	0.212 ± 0.040
Retrieve images (reports as input)		
Precision@5	0.233 ± 0.066	0.231 ± 0.059
Precision@10	0.248 ± 0.049	0.239 ± 0.041
Precision@50	0.247 ± 0.031	0.245 ± 0.026
mAP	0.159 ± 0.008	0.186 ± 0.007
Our SSL only		
Retrieve reports (image as input)		
Metric	Catania	Radboudumc
Precision@5	0.757 ± 0.021	0.768 ± 0.020
Precision@10	0.756 ± 0.019	0.768 ± 0.021
Precision@50	0.770 ± 0.009	0.775 ± 0.016
mAP	0.463 ± 0.017	0.583 ± 0.008
Retrieve images (reports as input)		
Precision@5	0.836 ± 0.030	0.850 ± 0.016
Precision@10	0.832 ± 0.028	0.851 ± 0.012
Precision@50	0.827 ± 0.012	0.845 ± 0.007
mAP	0.467 ± 0.010	0.580 ± 0.010
Classification + CLIP loss		
Retrieve reports (image as input)		
Metric	Catania	Radboudumc
Precision@5	0.780 ± 0.012	0.777 ± 0.016
Precision@10	0.779 ± 0.014	0.779 ± 0.018
Precision@50	0.789 ± 0.010	0.789 ± 0.017
mAP	0.546 ± 0.020	0.639 ± 0.018
Retrieve images (reports as input)		
Precision@5	0.847 ± 0.033	0.887 ± 0.016
Precision@10	0.847 ± 0.024	0.882 ± 0.017
Precision@50	0.846 ± 0.013	0.874 ± 0.014
mAP	0.515 ± 0.020	0.632 ± 0.022
Classification + our SSL loss		
Retrieve reports (image as input)		
Metric	Catania	Radboudumc
Precision@5	0.795 ± 0.030	0.794 ± 0.015
Precision@10	0.794 ± 0.028	0.796 ± 0.015
Precision@50	0.798 ± 0.020	0.802 ± 0.013
mAP	0.537 ± 0.023	0.632 ± 0.012
Retrieve images (reports as input)		
Precision@5	0.867 ± 0.030	0.880 ± 0.015
Precision@10	0.864 ± 0.024	0.877 ± 0.013
Precision@50	0.854 ± 0.021	0.873 ± 0.016
mAP	0.508 ± 0.019	0.620 ± 0.015

The performance is evaluated with precision@k (k values are 5, 10, 50) and mAP, reporting the average and the standard deviation (of the models involved in the k-fold cross-validation).

where the linking is effective are not involved in the classification. More details about the single concept linking are described in the Supplementary Material.

The visual ontology is shown in Fig. 6. For every concept, nine randomly selected patches are linked to the concept.

The upper part of Fig. 6 shows the linking between visual and textual concepts from the ExaMode ontology related to the ‘Positive Outcome’ (i.e. when findings are identified). The most relevant concepts in this part of the ontology are the ones related to the ‘Adenocarcinoma’, where it is possible to identify a small number of glands (that are not well defined).

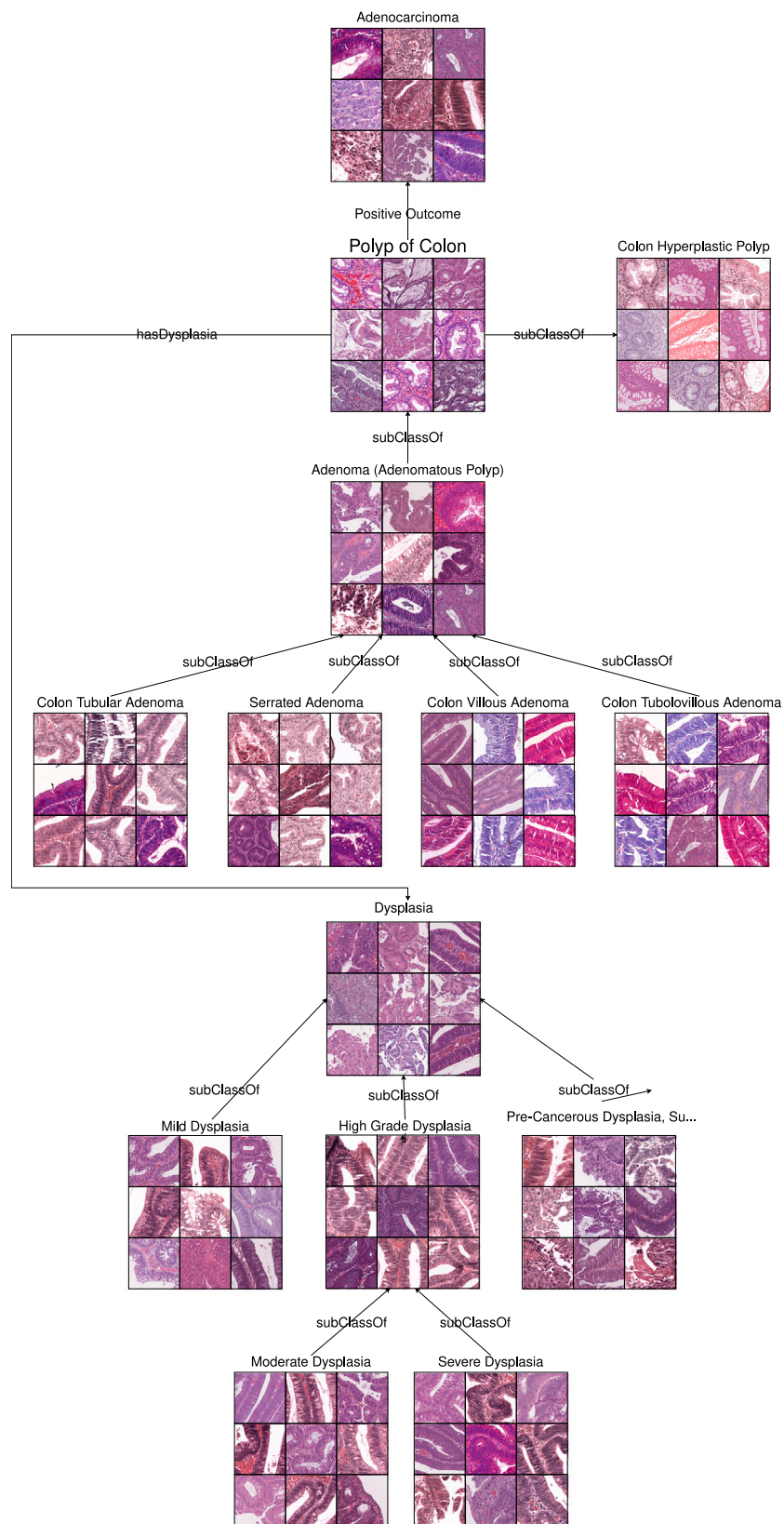


Fig. 6. The ExaMode visual ontology as a result of the linking between visual (patches) and textual concepts. The visual ontology includes concepts related to colon dysplasia concepts and on the polyp of colon.

**Table 6**

Overview of the results on the linking between visual and textual concepts, considering all training setups, at patch-level. For every concept, one hundred images are collected and reviewed by an expert pathologist. The model adopted for the linking is the one reaching the highest performance in terms of classification (when the classification term is involved) or retrieval (otherwise), among the training repetitions involved in the k-fold cross-validation.

Setup	True positives	Accuracy
CLIP loss only	204/1500	0.136
Our SSL loss only	747/1500	0.498
Classification + CLIP loss	745/1500	0.496
Classification + our SSL loss	911/1500	0.607

The central part of Fig. 6 shows the linking between visual and textual concepts from the ExaMode ontology related to colon polyps. Within the patches, it is possible to identify the presence of glands, better defined in terms of shape respect to the dysplasia patches, and the presence of a stroma less infiltrated.

The bottom part of Fig. 6 shows the linking between visual and textual concepts from the ExaMode ontology related to dysplasias. Within the patches, it is possible to identify the presence of deformed glands and of the stroma infiltrated by cells, usually related to the dysplasia condition. An important characteristic to underline is the fact that the patches linked to concepts are invariant to the color variations: the learned representation embeds features linked to tissue morphologies and to color variations.

Table 7 shows the results of the linking between visual and textual concepts, collected from reports, at WSI-level. The linking is effective, considering the number of concepts (15) and that the F1-score achieved by the multimodal architecture (trained combining classification and our SSL loss) is higher than the other setups. One aspect to stress is the fact that the models trained using only the CLIP loss function reach always 0, as a consequence of the low similarity values obtained with the concepts. This result can be explained considering the large amount of samples needed to train a deep learning model in a self-supervised fashion: in this only 6,000 samples were used, reducing the impact of the CLIP loss function. The result is relevant, because modality combination does not require any supervision and that most of the concepts where the linking is effective are not involved in the classification.

#### 4. Ablation study

The ablation study aims to investigate the contribution of the SSL loss functions on the classification of WSIs and of the image encoder in the multimodal architecture.

*The contribution of self-supervised loss functions on the classification of WSIs.* This section of the ablation study aims to investigate the contribution of the SSL loss functions on the WSI classification. The investigated loss functions are combinations of L1-loss, cosine similarity loss and NT-Xent loss; the application of CLIP loss.

The multimodal architecture is trained considering the classifications terms (for both WSIs and reports) and different combination of SSL losses and evaluated in terms of classification performance, considering the pathology workflow test partitions (Catania cohort, Radboudumc and their combination) and compared with the unimodal representation.

Table 8 shows an overview of the performance. The multimodal representation, learnt combining the three loss functions, reaches the highest performance in the Catania and the cumulative dataset, while instead the highest performance in the classification of Radboudumc data is reached by the multimodal (with CLIP loss function). Furthermore, comparing the performance with the unimodal representation, the multimodal architectures trained with our self-supervised loss function and with CLIP loss function are the only ones where the difference in performance is statistically significant (according to Wilcoxon Rank-Sum test).

*The contribution of the image encoder in the multimodal architecture.* This section of the ablation study aims to investigate the contribution of the image encoder on the WSI classification, considering four different MIL frameworks: ABMIL, CLAM, transMIL and ADMIL (the framework adopted in the paper). The multimodal architecture is trained considering two setups: the unimodal representation and the unimodal representation (with our self-supervised loss function) and evaluated in terms of classification performance, considering the pathology workflow test partitions (Catania cohort, Radboudumc and their combination).

Table 9 shows an overview of the performance, according to different MIL image encoders. The multimodal representation, learnt combining the three loss functions, reaches the highest performance in every test partition (Catania, Radboudumc and their combination), considering every MIL framework. However, the difference in performance is statistically significant for every partition only for the ABMIL and the ADMIL frameworks. The reason may be identified in the more complex architecture of CLAM and transMIL frameworks. These architectures implement mechanisms to identify relevant patches among the input bags. Therefore, the unimodal architecture, trained only with WSIs, reaches a plateau in the classification performance. Even if the classification performance is higher, we chose to adopt ADMIL: CLAM and transMIL framework are not trivial to set up and their tuning may not be trivial, requiring additional optimizations.

#### 5. Discussion

This paper presents a multimodal architecture, trained to combine histopathology images and textual reports to empower histopathology data representation. The multimodal architecture shows several advantages: a more robust representation of histopathology data, a solution to tackle data scarcity (still reaching accurate performance), and the possibility to link visual concepts from images to textual concepts from reports. The approach includes two input encoders, a CNN for images and a BERT model for textual reports, to process and combine modalities during the training. At testing time, the architecture works on single modalities. The architecture is trained on data collected from pathology workflows and evaluated on several tasks: first, WSI classification (on pathology workflow data, Section 2.4.1 and on publicly available data, Section 2.4.1); secondly, multimodal data retrieval (Section 2.4.1) and, lastly, the creation of multimodal ontologies, linking visual and textual concepts (Section 2.4.1).

The results achieved in the classification of images, both on pathology workflows and publicly available repositories, show that the multimodal data representation is more robust than the representation learnt adopting only images, leading to higher performance, and to better capability to generalize on unseen data. The multimodal architecture presented in the paper shows more robust characteristics for its application in the histopathology domain, compared with other Vision-Language Models presented in scientific literature to align the representations, such as CLIP (Radford et al., 2021), CoCa (Yu et al., 2022), ConVIRT (Zhang et al., 2020), PLIP (Huang et al., 2023), MI-Zero (Lu et al., 2023b), CONCH (Lu et al., 2023a). The main difference with these methods involves the amount of data adopted for training the architecture. While usually VLMs and SSL require a large amount of training samples, in this case only around 6,000 input samples are adopted: regarding the computational pathology domain, MI-Zero is trained with 33,000 pairs of image-text (over 5x bigger than this paper training dataset), PLIP on over 200,000 pairs (over 33x bigger than this paper training dataset), CONCH on over 1.17 million pairs (over 190x bigger than this paper training dataset); regarding computer vision domain, CLIP is trained on over 200,000 pairs (over 66,660x bigger than this paper training dataset). This detail is not trivial: the application of self-supervised loss functions, such as the one adopted to train the CLIP algorithm, does not lead to fine-grained histopathology data representations, required when training data are

**Table 7**

Overview of the results on the linking between visual and textual concepts, considering all training setups, at WSI-level. The performance is evaluated on the test partition collected from pathology workflow (Catania and Radboudumc). Being a multilabel problem, three threshold levels for the matching are proposed: 0.7, 0.8, 0.9. The performance is evaluated with weighted F1-score, reporting the average and the standard deviation (of the models involved in the k-fold cross-validation).

Setup	0.70		0.80		0.90	
	Catania	Radboudumc	Catania	Radboudumc	Catania	Radboudumc
CLIP loss only	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
Our SSL loss only	0.449 ± 0.055	0.485 ± 0.032	0.392 ± 0.068	0.432 ± 0.046	0.231 ± 0.047	0.173 ± 0.095
Classification + CLIP loss	0.462 ± 0.027	0.449 ± 0.040	0.419 ± 0.046	0.476 ± 0.057	0.269 ± 0.047	0.242 ± 0.084
Classification + Our SSL loss	0.484 ± 0.030	0.486 ± 0.051	0.435 ± 0.053	0.480 ± 0.040	0.286 ± 0.050	0.263 ± 0.114

**Table 8**

Overview of ablation study involving the self-supervised loss functions and their combinations. The loss functions are: L1-loss, cosine similarity loss, NT-Xent loss and CLIP loss. The performance is evaluated with weighted F1-score, reporting the average and the standard deviation (of the models involved in the k-fold cross-validation).

Loss function	Catania	Radboudumc	Cumulative
Unimodal	0.765 ± 0.013	0.771 ± 0.012	0.769 ± 0.011
L1	0.770 ± 0.015	0.775 ± 0.006	0.774 ± 0.008
Cosine similarity	0.773 ± 0.019	0.773 ± 0.010	0.774 ± 0.009
Contrastive	0.773 ± 0.009	0.778 ± 0.017	0.776 ± 0.012
L1 + Cosine similarity	0.770 ± 0.019	0.773 ± 0.015	0.772 ± 0.014
L1 + Contrastive	0.777 ± 0.016	0.785 ± 0.012*	0.783 ± 0.009
Cosine similarity + Contrastive	0.780 ± 0.010*	0.782 ± 0.010	0.782 ± 0.009*
CLIP loss function	0.785 ± 0.006*	0.782 ± 0.016*	0.784 ± 0.016*
<b>L1 + Cosine similarity + Contrastive</b>	<b>0.788 ± 0.011*</b>	<b>0.792 ± 0.012*</b>	<b>0.790 ± 0.009*</b>

\* The statistically significant results (compared with the unimodal representation, using the Wilcoxon Test) are reported.

**Table 9**

Overview of ablation study involving the image encoder backbone. The comparison involves four MIL frameworks, considering the unimodal and the multimodal representation (with our loss function as a self-supervised loss function) setups: ABMIL, CLAM, transMIL and ADMIL (the framework adopted in the paper). The performance is evaluated with weighted F1-score, reporting the average and the standard deviation (of the models involved in the k-fold cross-validation).

Image encoder setup	Catania	Radboudumc	Cumulative
ABMIL (Unimodal)	0.762 ± 0.012	0.769 ± 0.011	0.766 ± 0.010
ABMIL (Multimodal)	0.779 ± 0.006	0.793 ± 0.005*	0.787 ± 0.003*
CLAM (Unimodal)	0.778 ± 0.011	0.783 ± 0.017	0.781 ± 0.012
CLAM (Multimodal)	0.783 ± 0.013	0.798 ± 0.012*	0.792 ± 0.009*
transMIL (Unimodal)	0.783 ± 0.018	0.800 ± 0.015	0.793 ± 0.014
transMIL (Multimodal)	<b>0.803 ± 0.011*</b>	<b>0.806 ± 0.007</b>	<b>0.805 ± 0.006*</b>
ADMIL (Unimodal)	0.765 ± 0.013	0.771 ± 0.012	0.769 ± 0.011
ADMIL (Multimodal)	0.788 ± 0.011*	0.792 ± 0.012*	0.790 ± 0.009*

\* The statistically significant results (compared with the corresponding unimodal representation, using the Wilcoxon Test) are reported.

not large in size. For example, the multimodal architecture (trained adopting the CLIP setup) reaches the poorest performance in all the downstream tasks (i.e. multimodal retrieval and linking between visual and textual concepts). The reason may be identified in the design of CLIP, which requires two projection heads, that are l2-normalized, before aligning the multimodal representations. Due to this architectural design, the model risks overfitting on the relatively short amount of training data (compared with hundreds of thousands of samples required to train VLMs). The architecture design presented in the paper (a single projection head, shared among modalities, without any l2-normalization) allows a smoother modality alignment, that is reflected in the performance on every evaluated downstream task. Therefore, the loss function presented in the paper (a combination of L1-loss, cosine similarity loss and NT-Xent loss) better fits the scenario where multimodal datasets are not large in size, which is particularly true in the biomedical domain, therefore on the histopathology domain.

The benefits that the network shows to tackle the scarcity of training data are particularly clear when the classification performance is analyzed. The multimodal representation reaches the same performance as the unimodal representation, but using half of the training data (3,000 WSIs and reports vs. 6,000 WSIs). This result is remarkable

considering the application of the architecture in the histopathology domain (in general in the biomedical domain), where the collection of WSI annotations is time-consuming and not trivial. Therefore, the need for a reduced amount of training data to reach accurate and robust performance allows saving computational time, and reducing energy and carbon footprint for training deep learning models. The adoption of the multimodal architecture is also facilitated by the fact that usually WSIs (medical images in general) are paired and stored with the corresponding reports in the Laboratory Informative Systems, avoiding additional costs and efforts of collecting two medical modalities.

The multimodal representation generalizes better on unseen data, considering the results achieved on UNITOPatho and IMP-CRC data. In both cases, the image input branch of the architecture is trained with a brand new classifier, leaving the backbone of the architecture frozen, since data from publicly available datasets are annotated with different classes than the ones used to pre-train the network. The performance is compared with the unimodal representation (learnt only from WSIs). Multimodal representation reaches higher performance in both datasets (the difference is statistically significant), which is not trivial. However, the main implication of this result is the fact that the multimodal representation can be used as a valuable pre-trained backbone to train models on other classification tasks (for example including different classes) when dataset size is limited (UNITOPatho has around 200 WSIs in the training set, while IMP-CRC has around 800). Therefore, the pre-trained model can be easily fine-tuned with a small amount of data, guaranteeing good performance.

The results achieved in the multimodal data retrieval and on the linking of visual and textual concepts show that the multimodal data representation can be used to mine data and extract new knowledge from data. In both multimodal retrieval and concept matching tasks, the architecture shows robust performance, even if the architecture is not explicitly trained to solve those tasks. This feature of the network can be explained by considering the loss functions adopted to combine and align the image and report representations. In particular, the multimodal architecture shows competitive performance in the linking of visual and textual histopathology concepts at patch-level and WSI-level. The combination of SSL and weakly supervised learning enables the creation of visual ontologies for biomedical data, even when working with small training sets. This feature is crucial to understand in the

content, since it may pave the way to advancements in medicine and biomedical analysis domains, not limited to histopathology. Multimodal ontologies of biomedical data are still rare and difficult to create, often requiring human inputs. This fact not only prevents medical experts from properly benefitting from the recent opportunities offered by deep learning models, but it also prevents medical researchers in deep learning from benefitting from the opportunities offered by medical experts. Multimodal ontologies may allow medical experts to benefit from recent advancements in deep learning, considering aspects such as education, a richer data integration and standardization, and improved diagnostics. Biomedical visual ontologies can be adopted as educational tools for medical students, helping experts to understand the relationships among different structures and can improve their learning experience, also facilitating to better visualize information and peculiar conditions. Multimodal ontologies may help to standardize and better describe diseases, aligning medical terminology and data formats and contributing to enhance the exchanges among different healthcare systems, providing a shared framework for identifying diseases, symptoms and treatments, that can increase the agreement of medical experts on diagnosis. All these conditions would lead to enhanced diagnoses and better treatments provided by medical experts. On the other hand, multimodal ontologies may allow medical researchers in deep learning to benefit from the opportunities offered by medical experts, considering aspects such as data annotation and model refinement, accessible information, enhanced learning and prediction. The adoption of multimodal ontologies would allow to exploit large amounts of WSIs, currently unexploited and unannotated, to build weakly-supervised algorithms. The WSIs would be linked to relevant concepts, that could be used as weak labels. Furthermore, models can be refined on peculiar tissue structures that can be identified at patch-level. Therefore, by exploiting this feature, it is possible to build richer and richer datasets, including rare diseases. These datasets, built by linking high-level textual concepts to visual images, can be exploited to train new robust tools with relatively limited effort, starting a virtuous cycle not only in histopathology, but in the entire biomedical domain. Multimodal ontologies, linking visual and textual content, may allow to have more accessible and interpretable information, that can be exploited by several users, including clinicians and researchers. Also in this case, these aspects may dramatically help to design more robust algorithms to analyze biomedical data.

## 6. Conclusions

Linking visual and textual knowledge from biomedical data is still an unsolved task in biomedical domain, especially in domains where annotated datasets are few and heterogeneity is high, such as in computational pathology. This paper presents a multimodal architecture, including input branches for processing images and reports. The network is trained to classify images and reports, but due to the combination of modalities during training, it can be applied to solve other tasks, such as multimodal data retrieval and linking between visual and textual concepts. The multimodal architecture, trained combining both modalities, outperforms the same network, trained with only images, on the classification of pathology workflows (the Catania cohort and Radboudumc datasets) and publicly available data (two external datasets). Furthermore, the multimodal nature of the network also allows to retrieve multimodal data and to link textual concepts and images in a self-supervised fashion, providing a tool to mine large unlabeled datasets stored in hospital informative systems. In particular, the linking between visual and textual concepts allows to create a visual ontology of biomedical data. The application of biomedical ontology may dramatically lead to benefit for both medical experts and medical researchers. The multimodal approach can have a huge impact, not only on digital pathology but in general on biomedical sciences. Future works could target the development of multimodal representations

including different tissues, pathologies, concepts. Especially, linking together multimodal data representations of different domains may pave the way to unified multimodal representations of biomedical knowledge. The code and the multimodal architecture is publicly available on Github ([https://github.com/ilmaro8/multimodal\\_learning](https://github.com/ilmaro8/multimodal_learning)).

## CRediT authorship contribution statement

**Niccolò Marini:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Stefano Marchesin:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Marek Wodzinski:** Writing – review & editing, Writing – original draft, Formal analysis. **Alessandro Caputo:** Data curation. **Damian Podareanu:** Resources, Data curation. **Bryan Cardenas Guevara:** Resources, Data curation. **Svetla Boytcheva:** Writing – original draft, Methodology, Conceptualization. **Simona Vatrano:** Data curation. **Filippo Fraggetta:** Data curation. **Francesco Ciompi:** Methodology, Conceptualization. **Gianmaria Silvello:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Henning Müller:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Conceptualization. **Manfredo Atzori:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The authors do not have permission to share data.

## Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825292 (ExaMode, <http://www.examode.eu/>).

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.media.2024.103303>.

## References

- Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M.D., van der Laak, J., Bui, M.M., Vemuri, V.N., Parwani, A.V., Gibbs, J., Agosto-Arroyo, E., et al., 2019. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *J. Pathol.* 249 (3), 286–294.
- Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J., 2022. Multimodal biomedical AI. *Nature Med.* 28 (9), 1773–1784.
- Amal, S., Safarnejad, L., Omiye, J.A., Ghanzouri, I., Cabot, J.H., Ross, E.G., 2022. Use of multi-modal data and machine learning to improve cardiovascular disease care. *Front. Cardiovasc. Med.* 9.
- Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., MacWilliams, P., Mahdavi, S.S., Wulczyn, E., et al., 2022. Robust and efficient medical imaging with self-supervision. arXiv preprint arXiv:2205.09723.
- Barbano, C.A., Perlo, D., Tartaglione, E., Fiandrotti, A., Bertero, L., Cassoni, P., Grangotto, M., 2021. Unitopatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading. In: 2021 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 76–80.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901.
- Bulten, W., Kartasalo, K., Chen, P.-H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., et al., 2022. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Med.* 28 (1), 154–163.



- Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: fast and flexible image augmentations. *Information* 11 (2), 125.
- Campanella, G., Hanna, M.G., Geneslaw, L., Mirafior, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Med.* 25 (8), 1301–1309.
- Campanella, G., Kwan, R., Fluder, E., Zeng, J., Stock, A., Veremis, B., Polydorides, A.D., Hedvat, C., Schoenfeld, A., Vanderbilt, C., et al., 2023. Computational pathology at health system scale—self-supervised foundation models from three billion images. *arXiv preprint arXiv:2310.07033*.
- Carbonneau, M.-A., Cheplygina, V., Granger, E., Gagnon, G., 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit.* 77, 329–353.
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9650–9660.
- Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16144–16155.
- Chen, R.J., Krishnan, R.G., 2022. Self-supervised vision transformers learn visual concepts in histopathology. *arXiv preprint arXiv:2203.00585*.
- Chikontwe, P., Kim, M., Nam, S.J., Go, H., Park, S.H., 2020. Multiple instance learning with center embeddings for histopathology classification. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*. Springer, pp. 519–528.
- Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A., Courtiol, P., 2020. Self-supervision closes the gap between weak and strong supervision in histology. *arXiv preprint arXiv:2012.03583*.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Mac Kain, A., Saillard, C., Schiratti, J.-B., 2023. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv* 2023-07.
- Fraggetta, F., Garozzo, S., Zannoni, G.F., Pantanowitz, L., Rossi, E.D., 2017. Routine digital pathology workflow: the Catania experience. *J. Pathol. Inform.* 8 (1), 51.
- Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., et al., 2016. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* 70, 214–223.
- Goode, A., Gilbert, B., Harkes, J., Jukic, D., Satyanarayanan, M., 2013. OpenSlide: A vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.* 4 (1), 27.
- Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc. (HEALTH)* 3 (1), 1–23.
- Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B., 2009. Histopathological image analysis: A review. *IEEE Rev. Biomed. Eng.* 2, 147–171.
- Hanna, M.G., Reuter, V.E., Ardon, O., Kim, D., Sirintrapun, S.J., Schöffler, P.J., Busam, K.J., Sauter, J.L., Brogi, E., Tan, L.K., et al., 2020. Validation of a digital pathology system including remote review during the COVID-19 pandemic. *Mod. Pathol.* 33 (11), 2115–2127.
- Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I., 2020. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3852–3861.
- He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9729–9738.
- Heiliger, L., Sekuboyina, A., Menze, B., Egger, J., Kleesiek, J., 2022. Beyond medical imaging—a review of multimodal deep learning in radiology.
- Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J., 2023. A visual-language foundation model for pathology image analysis using medical Twitter. *Nature Med.* 29 (9), 2307–2316.
- Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P., 2020. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit. Med.* 3 (1), 136.
- Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning. In: *International Conference on Machine Learning*. PMLR, pp. 2127–2136.
- Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M., Madabhushi, A., 2019. HistoQC: an open-source quality control tool for digital pathology slides. *JCO Clin. Cancer Inform.* 3, 1–7.
- Javed, S.A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., Prakash, A., 2022. Additive mil: Intrinsically interpretable multiple instance learning for pathology. *Adv. Neural Inf. Process. Syst.* 35, 20689–20702.
- Junczyz-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Necker-mann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., et al., 2018. Marian: Fast neural machine translation in C++. *arXiv preprint arXiv:1804.00344*.
- Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N., 2021. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Trans. Med. Imaging* 40 (10), 2845–2856.
- Li, B., Li, Y., Eliceiri, K.W., 2021a. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14318–14328.
- Li, H., Yang, F., Zhao, Y., Xing, X., Zhang, J., Gao, M., Huang, J., Wang, L., Yao, J., 2021b. DT-MIL: deformable transformer for multi-instance learning on histopathological image. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. Springer, pp. 206–216.
- Liao, R., 2021. Multimodal Representation Learning for Medical Image Analysis (Ph.D. thesis). Massachusetts Institute of Technology.
- Liu, X., Cheng, H., He, P., Chen, W., Wang, Y., Poon, H., Gao, J., 2020. Adversarial training for large neural language models. *CoRR* abs/2004.08994.
- Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Zhang, A., Le, L.P., et al., 2023a. Towards a visual-language foundation model for computational pathology. *arXiv preprint arXiv:2307.12914*.
- Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.-S., Mahmood, F., 2023b. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19764–19775.
- Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* 5 (6), 555–570.
- Ma, E., 2019. NLP augmentation. <https://github.com/makcedward/nlpaug>.
- Marchesin, S., Giachelle, F., Marini, N., Atzori, M., Boytcheva, S., Buttafuoco, G., Ciompi, F., Di Nunzio, G.M., Fraggetta, F., Irrera, O., et al., 2022. Empowering digital pathology applications through explainable knowledge extraction tools. *J. Pathol. Inform.* 13, 100139.
- Marini, N., Atzori, M., Otálora, S., Marchand-Maillet, S., Müller, H., 2021a. H&E-adversarial network: a convolutional neural network to learn stain-invariant features through Hematoxylin & Eosin regression. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 601–610.
- Marini, N., Marchesin, S., Otálora, S., Wodzinski, M., Caputo, A., Van Rijthoven, M., Aswolinskiy, W., Bokhorst, J.-M., Podareanu, D., Petters, E., et al., 2022. Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations. *NPJ Digit. Med.* 5 (1), 102.
- Marini, N., Otálora, S., Podareanu, D., van Rijthoven, M., van der Laak, J., Ciompi, F., Müller, H., Atzori, M., 2021b. Multi-Scale-Tools: a python library to exploit multi-scale whole slide images. *Front. Comput. Sci.* 3, 684521.
- Marini, N., Otalora, S., Wodzinski, M., Tomassini, S., Dragoni, A.F., Marchand-Maillet, S., Morales, J.P.D., Duran-Lopez, L., Vatrano, S., Müller, H., et al., 2023. Data-driven color augmentation for H&E stained images in computational pathology. *J. Pathol. Inform.* 14, 100183.
- Menotti, L., Silvello, G., Atzori, M., Boytcheva, S., Ciompi, F., Di Nunzio, G.M., Fraggetta, F., Giachelle, F., Irrera, O., Marchesin, S., et al., 2023. Modelling digital health data: The ExaMode ontology for computational pathology. *J. Pathol. Inform.* 14, 100332.
- Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushirola, T., et al., 2017. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* 27 (Supplement\_III), S2–S8.
- Oliveira, S.P., Neto, P.C., Fraga, J., Montezuma, D., Monteiro, A., Monteiro, J., Ribeiro, L., Gonçalves, S., Pinto, I.M., Cardoso, J.S., 2021. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Sci. Rep.* 11 (1), 14358.
- Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Pallua, J., Brunner, A., Zelger, B., Schirmer, M., Haybaeck, J., 2020. The future of pathology is digital. *Pathol.-Res. Pract.* 216 (9), 153040.
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*. PMLR, pp. 8748–8763.
- Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al., 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Adv. Neural Inf. Process. Syst.* 34, 2136–2147.
- Srinidhi, C.L., Kim, S.W., Chen, F.-D., Martel, A.L., 2022. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Med. Image Anal.* 75, 102256.
- Stahlschmidt, S.R., Ulfenborg, B., Synnergren, J., 2022. Multimodal deep learning for biomedical data fusion: a review. *Brief. Bioinform.* 23 (2), bbab569.
- Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., Van Der Laak, J., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.* 58, 101544.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30.
- Veeranna, S.P., Nam, J., Mencia, E.L., Fürnkranz, J., 2016. Using semantic similarity for multi-label zero-shot classification of text documents. In: *Proceeding of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Elsevier, Bruges, Belgium, pp. 423–428.
- Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Mathieu, P., van Eck, A., Lee, D., Viret, J., et al., 2023. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*.
- Wang, J., Hu, X., Gan, Z., Yang, Z., Dai, X., Liu, Z., Lu, Y., Wang, L., 2021. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*.
- Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X., 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* 81, 102559.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al., 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Woolson, R.F., 2007. Wilcoxon signed-rank test. In: *Wiley Encyclopedia of Clinical Trials*. Wiley Online Library, pp. 1–3.
- Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y., 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.
- Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P., 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.
- Zhang, R., Zhang, Q., Liu, Y., Xin, H., Liu, Y., Wang, X., 2023. Multi-level multiple instance learning with transformer for whole slide image classification. *arXiv preprint arXiv:2306.05029*.