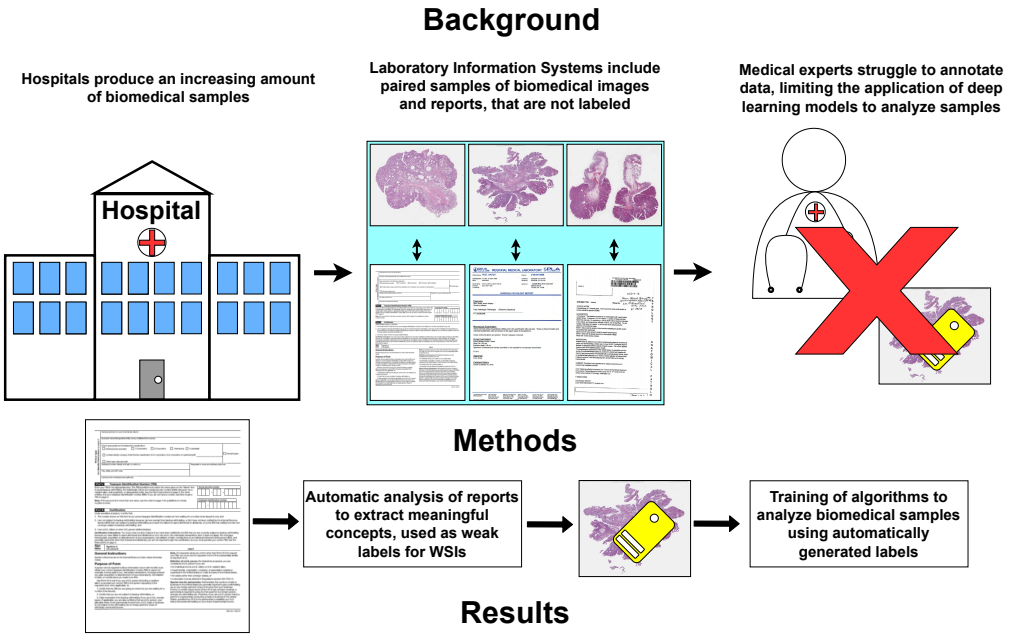
Graphical Abstract

Automatic Labels are as Effective as Manual Labels in Digital Pathology Images Classification with Deep Learning

Niccolò Marini*, Stefano Marchesin*, Lluis Borras Ferris, Simon Püttmann, Marek Wodzinski, Riccardo Fratti, Damian Podareanu, Alessandro Caputo, Svetla Boytcheva, Simona Vatrano, Filippo Fraggetta, Iris Nagtegaal, Gianmaria Silvello, Manfredo Atzori, Henning Müller



Highlights

Automatic Labels are as Effective as Manual Labels in Digital Pathology Images Classification with Deep Learning

Niccolò Marini*, Stefano Marchesin*, Lluis Borras Ferris, Simon Püttmann, Marek Wodzinski, Riccardo Fratti, Damian Podareanu, Alessandro Caputo, Svetla Boytcheva, Simona Vatrano, Filippo Fraggetta, Iris Nagtegaal, Gianmaria Silvello, Manfredo Atzori, Henning Müller

- Automatic weak labels can be adopted as weak labels to train deep learning algorithms to analyze biomedical samples, allowing the exploitation of large unannotated datasets.
- Algorithms generating weak labels with around 10% of mislabeled samples can be used to provide automatic weak labels.

Automatic Labels are as Effective as Manual Labels in Digital Pathology Images Classification with Deep Learning

Niccolò Marini*[a], Stefano Marchesin*[b], Lluis Borras Ferris[a], Simon Püttmann[c], Marek Wodzinski[a,d], Riccardo Fratti[a], Damian Podareanu[e], Alessandro Caputo[f,g], Svetla Boytcheva[h,i], Simona Vatrano[g], Filippo Fraggetta[g], Iris Nagtegaal[j], Gianmaria Silvello[b], Manfredo Atzori[a,k], Henning Müller[a,l]

[a]*Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais), Sierre, Switzerland,*
[b]*Department of Information Engineering, University of Padua, Padua, Italy,*
[c]*University of Applied Sciences and Arts Dortmund, Dortmund, Germany,*
[d]*Department of Measurement and Electronics, AGH University of Kraków, Krakow, Poland,*
[e]*SURFsara, Amsterdam, The Netherlands,*
[f]*Department of Pathology, Ruggi University Hospital, Salerno, Italy,*
[g]*Pathology Unit, Gravina Hospital Caltagirone ASP, Catania, Italy,*
[h]*Ontotext, Sofia, Bulgaria,*
[i]*Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria,*
[j]*Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands,*
[k]*Department of Neurosciences, University of Padua, Padua, Italy,*
[l]*Medical faculty, University of Geneva, 1211 Geneva, Switzerland,*

Abstract

The increasing availability of biomedical data is helping to design more robust deep learning (DL) algorithms to analyze biomedical samples. Currently, one of the main limitations to training DL algorithms to perform a specific task is the need for medical experts to label data. Automatic methods to label data exist; however, automatic labels can be noisy, and it is not completely clear when they can be adopted to train DL models. This

*Both authors contributed equally to this work. Corresponding author: Niccolò Marini.

paper aims to investigate under which circumstances automatic labels can be adopted to train a DL model on the classification of Whole Slide Images (WSI). The analysis involves multiple architectures, such as Convolutional Neural Networks (CNN) and Vision Transformer (ViT), and 10'604 WSIs as training partition, collected from three use cases: celiac disease, lung cancer, and colon cancer, which include respectively binary, multiclass and multilabel data. The results allow identifying 10% as the percentage of noisy labels that lead to train effective models for the classification of WSIs, reaching respectively F1-score 0.906, 0.757, 0.833. Therefore, an algorithm generating automatic labels needs to fit this criterion to be adopted. The application of the Semantic Knowledge Extractor Tool (SKET) algorithm to automatic extract concepts and use them as labels leads to performance comparable to that obtained with manual labels since it generates a percentage of noisy labels between 2% and 5%. Automatic labels are as effective as manual ones, achieving solid performance comparable to that obtained by training models with manual labels.

Keywords: Automatic Weak Labels, Deep Learning, Histopathology Image Classification, Noisy Labels,

---

1. Introduction

1.1. Background

Developing deep learning (DL) algorithms fosters the design of new tools that can be trained on clinical data without human intervention, especially in domains where annotations are expensive, such as histopathology. Histopathology is the gold standard to diagnose cancer (Van der Laak et al., 2021; De Matos et al., 2021). The domain involves the analysis of small tissue slices to identify microscopic findings related to dangerous diseases (Gurcan et al., 2009), such as cancer. Tissue slices undergo microscopic examination by a medical expert named a pathologist, who usually needs several minutes to analyze a single sample (Krupinski et al., 2013). Despite the increasing digitization of tissue samples, histopathological samples are rarely analyzed exploiting digital aid in clinical practice (Fraggetta et al., 2017, 2021). Digital pathology is a domain involving the management and digitization of tissue specimens, called Whole Slide Images (WSI). WSIs are high-resolution images stored with a pyramidal format, to capture different magnification levels of details (Merchant and Castleman, 2022). Usually, the

highest resolution levels result in a spatial high-resolution of 0.25–0.5μm per pixel, corresponding to an optical resolution of 20-40x. WSIs are usually coupled with pathology reports. Pathology reports are semi-structured free-text documents containing information about the patient's anamnesis, the tissue specimen type, and the findings and observations identified by a pathologist during the tissue examination (Hewer, 2020; Hanna et al., 2020). WSIs and reports are usually stored in the Laboratory Information System (LIS), which easily enables sample retrieval. The increasing collection of biomedical samples encourages the design of automatic tools to analyze WSIs under the computational pathology domain (Van der Laak et al., 2021; Madabhushi and Lee, 2016; Litjens et al., 2022). Most of the computational domain algorithms are currently based on deep learning, such as CNNs (Convolutional Neural Networks) or ViT (Visual Transformers) (Xu et al., 2023; Cifci et al., 2023).

Even if computational pathology algorithms show accurate and robust performance, in tasks such as WSI classification or segmentation, several challenges are still open, such as data labels (Madabhushi and Lee, 2016; Campanella et al., 2019; Van der Laak et al., 2021; Abels et al., 2019; Chen et al., 2022; Marini et al., 2024). Data labels are required to train supervised learning algorithms. However, the collection of labels is not trivial, considering both strong and weak annotations. Even if strong labels (i.e., pixel-wise annotations) usually achieve the most accurate performance when training a deep learning model, they require a pathologist to analyze samples, which can be time-consuming and often unfeasible (Karimi et al., 2020). Therefore, the research based on the analysis of WSIs is mostly based on the exploitation of weak (i.e., image-level) labels. Weak labels are related to the global image, even if they originate from a region of the image, including specific characteristics, such as cancer (Deng et al., 2020). Weak labels are inherently more noisy than pixel-wise annotations since the regions leading to a specific label may be a small percentage of the whole image (e.g., 1-2%). For this reason, algorithms based on weak labels require larger training datasets to reach accurate performance. Currently, most weakly-supervised algorithms in computational pathology are based on the Multiple Instance Learning (MIL) framework (Carbonneau et al., 2018), which models the whole image as a bag of instances, where only global annotations are available. MIL framework includes several algorithms, which lately showed high performance when adopted on large-scale datasets (Campanella et al., 2019; Ilse et al., 2018; Wang et al., 2019; Lu et al., 2021; Hashimoto et al.,

2020; Chen et al., 2022). For example, Campanella et al. (2019) showed that it is possible to reach almost perfect predictions on binary classification (cancer vs. non-cancer) using around 10'000 weakly annotated WSIs on three use cases: skin, breast, and prostate images. Weak labels are produced faster than strong ones, since they can be extracted from reports. For example, analyzing a report may take approximately 30 seconds/1 minute, compared with analyzing an image, which takes around an hour. However, human intervention is usually still required to analyze reports unless the Laboratory Information System (LIS), where the samples and corresponding reports are stored, has a specific structure to retrieve data automatically according to the characteristics that can be used as labels. Unfortunately, most LISs do not show this feature since they are organized in heterogeneous ways.

Automatic methods for extracting concepts from reports and using them as weak labels already exist (Marini et al., 2022), but noisy characteristics of weak labels can make automatic labeling ineffective. This paper investigates under which circumstances automatic labels (i.e., labels automatically generated by an algorithm) can be adopted to train deep learning models, alleviating the need for experts to annotate data. In particular, the goal is to identify when the results achieved using this type of label reach results comparable to those obtained using manual labels (i.e., labels produced by a medical expert) so that data included in LISs can be fully exploited to build more robust and accurate tools to diagnose diseases. The characteristics investigated in the paper involve the percentage of wrongly automatic labels necessary to reach comparable performance obtained with manual labels, the nature of labels (e.g., binary, multiclass, and multilabel), and the deep learning architecture (robust or less robust to noise). Wrongly automatic labels are annotations that are automatically produced by an algorithm and do not match the ground truth (i.e., they are manually made).

1.2. Contribution

The paper includes a comparison of deep learning architecture trained with automatic and manual labels on the classification of WSIs. The comparison involves two sets of experiments: a controlled scenario and a real-case scenario. In the controlled scenario, manual labels are randomly perturbed with different percentages of noise, simulating the output of an algorithm to generate automatic labels. The random perturbation involves modifying the labels. In the celiac disease use case, labels are flipped since the dataset includes binary annotations. A different class is assigned to a sample in the

lung cancer use case since the dataset includes multiclass annotations. In the colon cancer use case, the modifications involve one or more classes for every sample since the dataset includes multilabel annotations. In the real-case scenario, the Semantic Knowledge Extractor Tool (SKET) (Marchesin et al., 2022) is used to extract meaningful concepts from reports that are weak labels for the corresponding samples.

The analysis involves three tissue use cases, celiac disease, lung cancer, and colon cancer, composing a training dataset with over 10'000 WSIs, used to train three deep learning architectures: CLAM (Lu et al., 2021), transMIL (Shao et al., 2021) and Vision Transformer (ViT) (Chen et al., 2022).

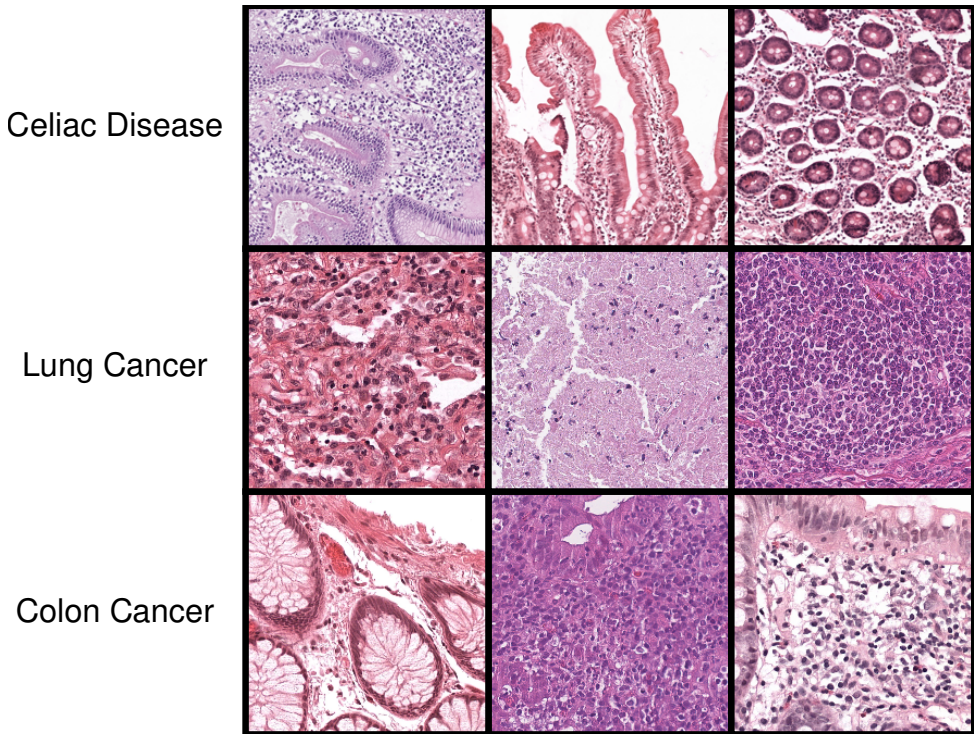


Figure 1: Overview of the tissue use cases analyzed in the paper. The upper line includes examples of duodenal tissue samples related to celiac disease. The central line includes examples of lung tissue samples. The bottom line includes examples of colon samples.

Celiac disease (CD) is an autoimmune disorder leading to damage in the small intestine, resulting in a range of gastrointestinal and systemic symptoms (Caio et al., 2019). Globally, celiac disease affects about 1-2% of the population (Lebwohl and Rubio-Tapia, 2021), with variations across regions.

In particular, the examination of biopsies aims to identify villous atrophy, crypt hyperplasia and increased intraepithelial lymphocytes. This paper labels duodenal samples with celiac disease or normal tissue (binary labels).

Lung cancer is the leading cause of death related to cancer worldwide (Schabath and Cote, 2019; Organization, 2023). It is categorized into two main primary groups: Non-Small Cell Lung Cancer (NSCLC), which represents the large majority of cases (about 85% of cases), and Small-Cell Lung Cancer (SCLC), which is less common, but more aggressive. Furthermore, NSCLC is further described with subtypes, such as LUng ADenocarcinoma (LUAD), LUng Squamous cell Carcinoma (LUSC). Diagnosis of lung cancer through biopsies often involves the identification of irregular cell patterns, architectural distortion, and increased cellular density (Travis, 2011). In this paper, lung samples are labeled with SCLC, LUAD, LUSC, and Normal Tissue.

Colon cancer is the fourth most often diagnosed cancer worldwide (Benson et al., 2018). Colon cancer diagnosis involves the identification of multiple concepts, such as the presence of cancer and the evaluation of polyp shapes and possible abnormalities leading to dysplasia. In this paper, colon samples are labeled with colon cancer, high-grade dysplasia (HGD), low-grade dysplasia (LGD), hyperplastic polyp and normal tissue (multilabel labels). Figure 1 shows some histopathological samples corresponding to the three tissues.

2. Materials and Methods

2.1. Dataset composition

The dataset used in this paper includes WSIs and reports (paired together) of celiac disease, lung cancer, and colon cancer collected from two hospitals: the Catania cohort and Radboudumc (RUMC).

WSIs are used to train and test different computer vision architectures on image-level classification. WSIs are gigapixel images, including tissue samples, that can exhibit significant heterogeneity, for example, in terms of staining (Marini et al., 2021a, 2023) and sample types. Image heterogeneity is a consequence of different acquisition procedures across laboratories related to the chemical reagents applied to the specimen and to the slide scanners as a whole. One of the main consequences of the heterogeneity is the stain variability, leading to different color variations, intensity, and uniformity of stains across different slides (as shown in Figure 1). The WSIs collected in

this dataset also show the same characteristics, aiming to replicate a common scenario in digital pathology. WSIs collected from the Catania cohort were scanned with two 3DHistech scanners and two Aperio scanners and stored with a magnification of 20-40x; WSIs collected from RUMC were scanned using 3DHistech scanners, mainly stored at 40x magnification.

Reports are used to extract meaningful concepts used as weak automatic labels to train the model to classify WSIs. Reports include free-text descriptions summarizing the findings from tissue examination. The findings are reported in a field named 'Conclusion', containing either macroscopic or microscopic observations. Even if a report includes many fields, only the findings are relevant for the analysis proposed in the paper. Therefore, additional patient information, such as family history or personal data, is discarded. Textual reports show heterogeneity, mainly related to the source language and the textual content. Reports are collected from an Italian and a Dutch hospital, therefore they have to be translated into English, to standardize the analysis. The textual content slightly differs across sources because the Catania cohort reports are related to a single slide, while the RUMC reports include a specific field for the findings identified in a tissue block, which may encompass multiple slides related to different images. The textual content slightly differs across sources because the Catania cohort reports contain a field specifically for the findings identified in a single slide, while the RUMC reports include a field specifically for the findings identified in a tissue block, which may encompass multiple slides. Therefore, RUMC reports needed a pre-processing step to separate the content and link it to the corresponding WSI. Furthermore, samples are collected over the ears and produced by many different pathologists, each adopting a unique style of writing.

The dataset includes samples collected from three different use cases: celiac disease, lung cancer, and colon cancer. Data are randomly selected from LISs to simulate a real-case scenario. The goal is to show that the approach can be generalized to different types of tissue (both in terms of images and reports). Different labels are used: celiac disease samples are annotated with binary labels, lung samples with multiclass labels, and colon samples with multilabel samples.

Table 1 includes a detailed composition of data related to celiac disease collected from pathology reports, split into training and testing partitions. Data are labeled with binary labels: celiac disease and normal tissue.

Table 2 includes a detailed composition of data related to lung cancer collected from pathology reports, split in training and testing partitions.

Table 1: Composition of the samples related to the celiac disease use case, considering automatically generated labels (automatic labels) and ground truth labels (manual labels). Data are labeled with binary labels: celiac disease and normal tissue. The dataset is split into training and testing partitions. The model is trained and validated, adopting a 10-fold cross-validation approach.

Source	**Celiac Disease**	**Normal Tissue**	**Total**
Training dataset: Automatic Labels			
Catania	47	711	758
RUMC	217	524	741
Total	264	1235	1499
Training dataset: Manual Labels			
Catania	61	697	758
RUMC	223	518	741
Total	284	1235	1499
Testing dataset			
Catania	10	83	93
RUMC	37	63	100
Total	47	146	193

Table 2: Composition of the samples related to the lung cancer use case, considering automatically generated labels (automatic labels) and ground truth labels (manual labels). Data are labeled with multiclass labels: Small-Cell Cancer, Non-Small Adenocarcinoma Cell Cancer, Non-Small Squamous Cell Cancer, Normal Tissue. The dataset is split into training and testing partitions. The model is trained and validated, adopting a 10-fold cross-validation approach.

Source	**SCLC**	**LUAD**	**LUSC**	**Normal**	**Total**
Training dataset: Automatic Labels					
Catania	49	526	250	226	1051
RUMC	1	262	195	1041	1499
Total	50	788	445	1267	2550
Training dataset: Manual Labels					
Catania	50	519	271	211	1051
RUMC	1	260	173	1065	1499
Total	51	779	444	1276	2550
Testing dataset					
Catania	12	62	67	32	173
RUMC	0	55	29	110	194
Total	12	117	96	142	367

Data are labeled with multiclass labels: Small-Cell Cancer, Non-Small Adenocarcinoma Cell Cancer, Non-Small Squamous Cell Cancer, Normal Tissue.

Table 3: Composition of the samples related to the colon cancer use case, considering automatically generated labels (automatic labels) and ground truth labels (manual labels). Data are labeled with multilabel annotations: Adenocarcinoma, High-Grade Dysplasia (HGD), Low-Grade Dysplasia (LGD), Hyperplastic Polyp, Normal Tissue. Due to the multilabel nature of labels, the total samples for each class may not correspond to the total number of samples. The dataset is split into training and testing partitions. The model is trained and validated adopting a 10-fold cross-validation approach.

Source	Adenocarcinoma	HGD	LGD	Hyperplastic	Normal	Total
Training dataset: Automatic Labels						
Catania	776	761	1288	511	596	3095
RUMC	383	377	853	943	1341	3460
Total	1159	1138	2141	1454	1937	6555
Training dataset: Manual Labels						
Catania	865	774	1273	535	570	3095
RUMC	394	362	878	965	1309	3460
Total	1259	1136	2151	1500	1879	6555
Testing dataset						
Catania	111	96	113	32	98	348
RUMC	75	65	146	119	193	520
Total	186	161	259	151	291	868

Table 3 includes a detailed composition of data related to colon cancer collected from pathology reports, split into training and testing partitions. Data are labeled with multilabel labels: Adenocarcinoma, High-Grade Dysplasia (HGD), Low-Grade Dysplasia (LGD), Hyperplastic Polyp, Normal Tissue.

2.2. Data analysis pipeline

The training schema is based on computer vision algorithms to classify WSIs, comparing the performance of automatic and manual labels during the training. Those algorithms are based on weak labels since they are easier to collect, even if they still require the intervention of medical experts. This paper adopts three different MIL backbones: two CNNs, CLAM and transMIL, and a ViT. The architectures are trained to evaluate the effect that automatic labels may have on the training of models to classify WSIs. Firstly, they are trained with noisy labels, randomly generated to perturb the manual labels with a different percentage (1,2,5,10,20,50%) of noise. This experiment's goal is to evaluate the effect that noisy labels have on a model's performance. However, this setup does not fit a real-case scenario where automatic labels are adopted. Noisy labels may be considered wrongly labeled samples, but not all label mistakes are equally likely to occur. Consider, for

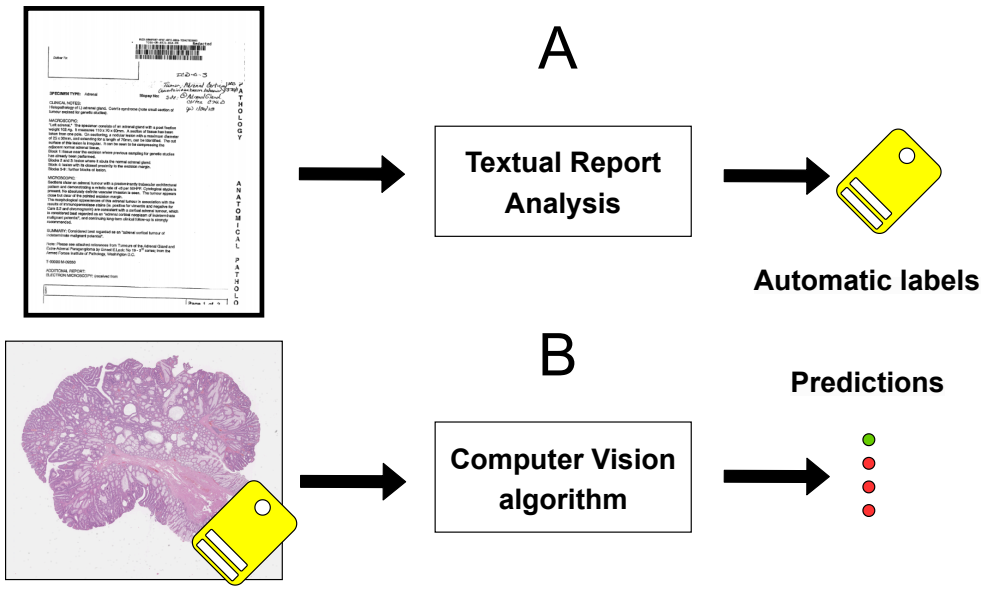


Figure 2: Overview of the data analysis pipeline proposed in the paper. It includes two steps. The first step (A) involves the analysis of textual reports to extract meaningful concepts that can be used as weak (automatic) labels for WSIs. The second step (B) involves image analysis through computer vision algorithms that are transparent to the user and can be exchanged to predict the content of the images.

example, weak labels inferred by reports: some reports, due to their content, may be more easily mislabeled. For this reason, a real tool to extract concepts from reports is adopted: the Semantic Knowledge Extractor Tool (SKET) (Marchesin et al., 2022). The goal of its application is to evaluate a real-world scenario in which a tool to generate automatic labels is adopted. The adoption of SKET allows to have samples that are not randomly mislabeled, but rather mislabeled due to the content of the corresponding reports, that can be hard to interpret. This condition also helps test the rules on the percentage of mislabeled samples identified with randomly perturbated samples.

Figure 2 shows an overview of the data analysis pipeline.

2.3. Computer vision architectures

The paper compares three computer vision algorithms to classify WSIs as backbones to evaluate the effect of noisy labels on different architectures,

including two CNNs and a ViT. The CNNs have a ResNet34 backbone, while the ViT has a backbone similar to the one shown in Chen et al. (2022), considering a single magnification level. In both cases, the backbones are designed to output an embedding of size 128 representing a single WSI, so that the same classifier can be adopted for all architectures, modifying the output classes based on the use case.

CLAM. Clustering-constrained Attention Multiple Instance Learning (CLAM) (Lu et al., 2021) is a MIL framework based on an attention-based network that highlights relevant regions inside the WSI to improve the WSI-level prediction. CLAM exploits a mechanism on the single instances to aggregate them on clusters, according to the instance similarity, to enrich the WSI representation and reach higher WSI-level predictions. CLAM can have one or more attention branches, depending on the number of classes. In this paper, a single attention branch (CLAM_SB) is used when the model is used on celiac disease (binary labels), while instead a multiple attention branch (CLAM_MB) is used on the other two use cases.

transMIL. transMIL (Shao et al., 2021) is a MIL framework developed to exploit the morphological and spatial characteristics of WSIs. Even if morphological and spatial characteristics of images are important, the attention mechanism does not consider them when evaluating input instances. transMIL exploits Transformer architectures (Vaswani et al., 2017) to highlight relationships between single instances, modeling input instances as a sequence of tokens and evaluating the similarity among instances.

Vision Transformer. Vision Transformer (Sharir et al., 2021; Han et al., 2020) is a deep learning architecture adopted to analyze images, adopting the self-attention mechanism to process input data instead of convolutional layers, showing more competitive performance in terms of accuracy and efficiency. The architecture processes input data as a sequence of input tokens that are small sub-regions of the input image (usually 16x16 pixels). The architecture includes 12 encoder layers producing the embedding to feed the classifier.

2.4. Semantic Knowledge Extractor Tool (SKET)

SKET (Marchesin et al., 2022) is an unsupervised algorithm combining a rule-based expert system with machine learning models, chosen to extract meaningful concepts from reports and use them as weak labels for

WSIs (Marchesin et al., 2022; Menotti et al., 2023). The algorithm includes Named Entity Recognition, Entity Linking and Data Labeling. Named Entity Recognition involves pre-trained models (ScispaCy models (Neumann et al., 2019)), developed to work on biomedical data, and large Word2Vec word vectors (Mikolov et al., 2013) trained on the PubMed Central Open Access Subset (Mikolov et al., 2013). Entity Linking combines similarity-matching techniques to match ad-hoc concepts to a reference ontology. Data Labeling involves mapping the concepts with a set of annotation classes. SKET is an unsupervised model; therefore, no training data are required to tune it. This feature is relevant because it does not require data annotation for training, unlike other Natural Language Processing (NLP) algorithms.

3. Experimental Setup

3.1. Image pre-processing

Image pre-processing includes the WSI splitting into patches. Because of their gigapixel characteristics, WSIs usually do not fit modern GPU hardware memory; therefore, they have to be split into patches. In this paper, WSIs are split 224x224 pixel patches using the Multi_Scale_Tools library (Marini et al., 2021b). The choice of the size is related to the characteristics of ResNet34 backbone, requiring fixed input size. Patches are extracted from magnification 5x, considering celiac samples, while lung and colon patches are sampled from magnification 10x. The magnifications are chosen considering that the magnification allows the identification of peculiar morphological features, which are useful for the classification task.The choice of the magnification to examine is driven by the characteristics of the problem to solve: celiac disease diagnosis requires to identify the villous shape and the crypts, therefore 5x magnification is chosen; on the other hand, lung and colon require a more refined level of magnification, because the shape of glands is as relevant as the cell infiltration, therefore 10x is chosen. Not all sampled patches are selected: the ones from background regions are discarded, being not informative. Identifying background regions involves applying HistoQC tool (Janowczyk et al., 2019), which generates tissue masks.

3.2. Report pre-processing

The report pre-processing only involves their translation into English. Original reports are stored in Italian and Dutch, depending on the workflow from which they are collected. The translation is necessary because

state-of-the-art NLP algorithms are mostly developed to work with inputs in English. MarianMT neural machine translation models (Junczys-Dowmunt et al., 2018) are used to translate the content of the reports to English.

3.3. Architecture pre-training

The backbones of deep learning algorithms to analyze images are pre-trained using self-supervised algorithms: simCLR (Chen et al., 2020) for the CNNs (CLAM and transMIL), DINO v2 (Oquab et al., 2023) for the ViT.

Both algorithms are adopted to learn meaningful features from unannotated input data, exploiting similarities and dissimilarities between input samples. In this paper, the input data for the algorithms are the patches sampled from the training partition. Since data are unannotated, no information is available regarding patch similarity. Therefore, data augmentation is adopted: samples are similar to their augmented versions and dissimilar from the other samples within a batch. The algorithms differ in the data augmentation strategy. simCLR is designed for CNNs and its augmentation pipeline includes several operations, applied with a probability of 0.5: random rotations (90/180/270 degrees), vertical/horizontal flipping, hue-saturation-contrast (HUE) color augmentation, RGB shift, color jitter, gaussian noise, elastic transformation, grid distortions. DINO is designed for ViT and involves a knowledge distillation mechanism: two networks, a teacher and a student, are involved in the training. The teacher is a larger model producing outputs that the student aims to mimic and replicate. Both models are directly trained with two different augmented versions of input samples. However, the student is also trained with a cropped version (96x96 pixels) of the teacher inputs. The DINO v2 augmentation pipeline includes two pipelines: the first one includes color jitter, horizontal/vertical flipping, Gaussian blur, and solarization.

3.4. Image data augmentation pipeline

Augmentation library (Buslaev et al., 2020) is adopted to apply data augmentation to input images. The operations involved are random rotations (90/180/270 degrees), vertical/horizontal flipping and hue-saturation-contrast (HUE) color augmentation. The operations from the data augmentation pipeline are selected with a probability of 0.5 and applied at image-level, so that all the patches are augmented consistently.

3.5. Metric to evaluate the performance

The performance of the models is evaluated in terms of WSI classification using the weighted F1-score. The classification problem can be defined as binary (celiac disease), multiclass (lung cancer), or multilabel (colon cancer). F1-score is a metric used to measure the accuracy of a classifier, combining recall and precision. Precision evaluates how well a classifier is robust to avoid predicting negative samples as positive ones, while recall evaluates how well it correctly classifies all the positive samples. Data may show unbalanced class distribution in all the use cases, since they are randomly selected from workflows, aiming to simulate a real-case scenario. For this reason, a weighted macro F1-score is adopted Weighted F1-score tackles class imbalance, evaluating the F1-scores for the single classes and then averaging them according to the class support (number of true samples for the class). The weighted F1 Score is reported as the average and standard deviation of the ten experiment repetitions evaluated on the test partition.

3.6. Statistical significance test

The performance difference among different setups is evaluated through the Wilcoxon Rank-Sum test (Woolson, 2007). The test aims to establish if the results of two different experiments are statistically significantly different (p-value ¡ 0.05).

3.7. K-fold cross-validation

All the setups presented in the paper are trained using k-fold cross-validation to evaluate the model's robustness to the data used for training. The training partition is divided into k folders (k=10 in this paper). During every training repetition, k-1 folders are used to train the model, while the other group is used to validate it. Data are split into partitions considering the patients so that WSIs collected from a patient cannot be in two different partitions

3.8. Hardware and Software

The experiments are developed exploiting Python libraries. The deep learning algorithms are implemented and trained using PyTorch 2.2.0 and run on a Tesla V100 GPU. WSIs are accessed using openslide 3.4.1 (Goode et al., 2013). WSI pre-processing involves Multi_Scale_Tools library (Marini et al., 2021b) and data augmentation is applied using albumentations 1.3.1 (Buslaev et al., 2020). The performance of the model is quantitatively evaluated using the metrics implemented by sci-kit-learn 0.22.

3.9. Hyperparameters

The optimal configuration setup of both CNN and ViT hyperparameters is identified using the grid search algorithm. Considering the validation partition, the optimal set reaches the lowest loss function of the classification of WSIs. The parameters tested with the grid search algorithm are: the batch size (4 selected; 1,2,4,8 tested); the CNN optimizer (Adam selected); the ViT optimizer (Adam selected; Adam, LARS and AdamW tested); the number of epochs when the CNN model is trained (15; over this number of epochs, the loss function evaluated on the validation partition no longer decreases); the number of epochs when the HIPT model is trained (15; over this number of epochs, the loss function evaluated on the validation partition no longer decreases); the learning rate (10^{-4}; 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} were tested); the decay rate (10^{-4}; 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} were tested); the number of nodes in the intermediate layer after the ResNet and the ViT backbone (128; 64, 128, 256, 512 were tested).

4. Results

4.1. Automatic labels

Table 4: Overview of SKET's performance on extracting meaningful concepts from pathology reports, evaluated in terms of F1-score. The performance is evaluated comparing the concepts extracted by SKET as labels and the ground truth labels. The algorithm is evaluated considering the training partitions of three use cases (celiac disease, lung cancer, colon cancer), since SKET requires no training. The results are assessed based on data from Catania and RUMC and their combination for every tissue use case.

Use case	Catania	RUMC	Cumulative
Celiac Disease	0.860	0.964	0.944
Lung Cancer	0.969	0.975	0.976
Colon Cancer	0.976	0.961	0.971

Meaningful concepts can be extracted from pathology reports without the need for human intervention and can be adopted as weak labels, dramatically reducing the time needed to collect labels.

The performance of SKET (a tool to extract weak labels from reports) is evaluated on the training partition of the three use cases since SKET is a ruled-based algorithm that does not require any training. The extracted concepts are compared with the manual labels provided by medical experts.

Table 4 summarizes the results. SKET reaches a weighted F1-score over 0.944 on every use case, considering the cumulative testing partition. On the single pathology workflows, the lowest performance is reached considering the Catania testing partition of celiac disease data (0.860). Otherwise, the algorithm reaches high-level performance, always over 0.960 in terms of F1-score.

Table 5: Overview of the time needed by SKET and a human expert to annotate reports. The comparison involves three possible durations for a human expert and two for SKET. The values chosen for a human expert are 1s, 10s, and 30s, respectively, extremely (but unfeasible) fast annotators, ceiling of the annotation range, and floor of the annotation range. The values chosen for SKET are 0.006s and 0.03s, respectively, with the ceiling of the annotation range and the floor of the annotation range. The comparison is made considering 10'000 annotated reports and includes the percentage of time saved.

Time per iteration	Automatic (mins)	Manual (mins)	Percentage saved
A: 0.006s / M: 1s	1	166,6666667	99,40 %
A: 0.03 / M: 1s	5	166,6666667	97,00 %
A: 0.006s / M: 10s	1	1666,666667	99,94 %
A: 0.03 / M: 10s	5	1666,666667	99,70 %
A: 0.006s / M: 30s	1	5000	99,98 %
A: 0.03 / M: 30s	5	5000	99,90 %

Effectively, SKET can be adopted to mine unlabeled datasets and annotate large amounts of data, which can be used to train deep learning models. Table 5 summarizes the results. When tested on a Tesla V100 GPU, SKET requires between 0.006 (ceiling annotation time) and 0.03 (floor annotation time) seconds to extract concepts from a report, depending on its length. A human expert needs between 10s (ceiling annotation time) and 30s (floor annotation time) to extract concepts from a report, depending on its length and content. Considering the worst-case scenario for SKET and the best-case scenario for a human expert (0.03s vs 10s), the algorithm is still around 333 times (0.03 / 10) faster than a human. For instance, in the best-case scenario, the weak labeling of 10,000 WSIs would require 300,000 seconds (around 83 hours, without breaks) for human experts; in the worst-case scenario, it would require 300 seconds (five minutes) to SKET. Therefore, the application of SKET leads to save 99.7% of time required in comparison with human experts. Even considering an unreasonable effectiveness of a human experts, such as 1s per iteration, would lead to save 97% of the time. A detail

relevant to stress is that the comparison considers the best possible condition for a human expert (no breaks, no wasted time, ceiling performance) and the worst condition for SKET (floor performance).

4.2. Celiac disease

The classification performance of multiple computer vision architectures trained with binary automatically annotated data to classify celiac disease WSIs is as effective as the performance reached by models using manually annotated data.

Tables 6 and 7 summarize the results. The highest performance using manual labels is reached using a ViT architecture (F1-score = 0.914 ± 0.014 on the test partition), even if on the Catania partition transMIL shows the highest performance. The results are still similar for the three architectures.

Table 6 shows the classification performance obtained using binary manual labels and noisy labels. This experiment aims to investigate general rules for the adoption of automatic labels on the binary classification of WSIs. Considering all the architectures, the performance is similar to the one obtained using manual labels, especially until 10% of training samples are wrongly annotated. The difference in terms of performance is not statistically significant. When the percentage of wrongly-annotated training is 20% (or more) the performance degrades and the difference, compared with manual labels, is statistically significant, suggesting this percentage of wrongly-annotated labels can be considered as a threshold for the adopting of automatic weak labels in a binary classification scenario.

Table 7 compares automatic labels generated with SKET and manual labels. The comparison among automatic and manual labels shows a F1-score equal to 0.944, suggesting that the algorithm should lead to performance similar to the one obtained with noisy labels when the percentage of mislabeled data is between 2% and 5%. The results confirm the hypothesis since the performance is slightly worse than the one obtained using manual labels, but the gap is not statistically significant (according to the Wilcoxon Rank-Sum test, comparing every setup to the one where manual labels are used), showing the effectiveness of automatic labels in a binary classification scenario.

4.3. Lung cancer

The classification performance of multiple computer vision architectures trained with multiclass automatically annotated data to classify lung cancer

Table 6: Results on the classification of celiac disease, in terms of F1-score. The performance is evaluated considering three computer vision architectures: CLAM, transMIL, ViT. The architectures are trained with manual weak binary labels and with noisy, weak labels, randomly perturbated according to different percentages of noise. The percentage of noisy labels is reported in the 'Noisy Labels' column, while the accuracy of the labels is reported in terms of F1-score, 'F1 labels' column. The goal is to evaluate the effect that noisy weak labels have on the binary classification of WSIs. For every setup, the F1-score average and standard deviation of the classification performance are reported, considering the models trained with the 10-fold cross-validation. The setups where the difference is statistically significant in terms of performance (compared with the models trained with manual labels) are marked with an asterisk (*).

Noisy Labels	**F1 Labels**	**Model**	**Catania**	**RUMC**	**Cumulative**
Manual	-	CLAM_SB	0.958 ± 0.009	0.846 ± 0.023	0.900 ± 0.012
		transMIL	0.968 ± 0.009	0.850 ± 0.019	0.906 ± 0.010
		ViT	0.953 ± 0.011	0.877 ± 0.021	0.914 ± 0.014
1%	0.977	CLAM_SB	0.954 ± 0.016	0.849 ± 0.024	0.900 ± 0.018
		transMIL	0.968 ± 0.009	0.864 ± 0.010	0.914 ± 0.007
		ViT	0.954 ± 0.014	0.896 ± 0.019	0.925 ± 0.010
2%	0.968	CLAM_SB	0.951 ± 0.012	0.873 ± 0.021	0.911 ± 0.014
		transMIL	0.965 ± 0.011	0.853 ± 0.021	0.907 ± 0.010
		ViT	0.944 ± 0.017	0.877 ± 0.021	0.910 ± 0.013
5%	0.933	CLAM_SB	0.951 ± 0.019	0.862 ± 0.019	0.905 ± 0.017
		transMIL	0.958 ± 0.012*	0.857 ± 0.018	0.905 ± 0.011
		ViT	0.938 ± 0.026	0.880 ± 0.026	0.910 ± 0.020
10%	0.909	CLAM_SB	0.952 ± 0.013	0.862 ± 0.023	0.905 ± 0.017
		transMIL	0.953 ± 0.026*	0.838 ± 0.033	0.893 ± 0.027
		ViT	0.957 ± 0.014	0.860 ± 0.023	0.906 ± 0.014
20%	0.804	CLAM_SB	0.922 ± 0.026*	0.819 ± 0.029	0.869 ± 0.023*
		transMIL	0.933 ± 0.024*	0.822 ± 0.013*	0.875 ± 0.016*
		ViT	0.925 ± 0.017*	0.834 ± 0.025*	0.879 ± 0.017*
50%	0.566	CLAM_SB	0.537 ± 0.228*	0.450 ± 0.081*	0.490 ± 0.145*
		transMIL	0.765* ± 0.097*	0.502* ± 0.02*	0.633 ± 0.041*
		ViT	0.440 ± 0.302*	0.459 ± 0.029*	0.480 ± 0.141*

WSIs is as effective as the performance reached by models using manually annotated data.

Tables 8 and 9 summarize the results. The highest performance using manual labels is reached using a ViT architecture (F1-score = 0.763 ± 0.012) on both test partitions, dramatically outperforming the other two architectures (CLAM reaches 0.674 ± 0.016, while transMIL reaches 0.696 ± 0.016).

Table 8 shows the classification performance obtained using multiclass manual labels and noisy labels. This experiment aims to investigate general

Table 7: Results on the classification of celiac disease, in terms of F1-score. The performance is evaluated considering three computer vision architectures: CLAM, transMIL, ViT. The architectures are trained with automatic and manual weak binary labels generated by extracting meaningful concepts from the corresponding pathology report using the SKET algorithm. The performance of SKET is reported in the 'Noisy labels' column. The goal is to evaluate the effectiveness of automatic labels on the binary classification of WSIs. For every setup, the F1-score average and standard deviation of the classification performance are reported, considering the models trained with the 10-fold cross-validation. The setups where the difference is statistically significant in terms of performance (compared with the models trained with manual labels) are marked with an asterisk (*).

Noisy Labels	**F1 Labels**	**Model**	**Catania**	**RUMC**	**Cumulative**
Automatic	0.944	CLAM_SB	0.948 ± 0.015	0.857 ± 0.017	0.901 ± 0.013
		transMIL	0.960 ± 0.012	0.845 ± 0.017	0.900 ± 0.014
		ViT	0.938 ± 0.023	0.889 ± 0.024	0.915 ± 0.015
Manual	-	CLAM_SB	0.958 ± 0.009	0.846 ± 0.023	0.900 ± 0.012
		transMIL	0.968 ± 0.009	0.85 ± 0.019	0.906 ± 0.010
		ViT	0.953 ± 0.011	0.877 ± 0.021	0.914 ± 0.014

rules for the adoption of automatic labels on the multiclass classification of WSIs. Considering all the architectures, the performance is similar to the one obtained using manual labels, especially until 20% of training samples are wrongly-annotated, the difference in terms of performance is not statistically significant. When the percentage of wrongly-annotated training is 50% the performance degrades and the difference, compared with manual labels, is statistically significant, suggesting this percentage of wrongly annotated labels can be considered as a threshold for the adoption of automatic weak labels in a multiclass classification scenario.

Table 9 includes the comparison of automatic labels and manual labels. This comparison represents a real-case scenario of automatic data labeling, where automatic labels are generated by extracting concepts from reports. The comparison among labels shows a F1-score equal to 0.976, suggesting that the algorithm should lead to performance similar to the one obtained in the previous experiment using 2% and 5%. The results confirm the hypothesis since the performance is slightly worse than the one obtained using manual labels, but the gap is not statistically significant (according to the Wilcoxon Rank-Sum test, comparing every setup to the one where manual labels are used).

Table 8: Results on the classification of lung cancer, in terms of F1-score. The performance is evaluated considering three computer vision architectures: CLAM, transMIL, ViT. The architectures are trained with manual weak multiclass labels and with noisy weak labels, randomly perturbated according to different percentages of noise. The percentage of noisy labels is reported in the 'Noisy Labels' column, while the accuracy of the labels is reported in terms of F1-score, 'F1 labels' column. The goal is to evaluate the effect that noisy weak labels have on the multiclass classification of WSIs. For every setup, the F1-score average and standard deviation of the classification performance are reported, considering the models trained with the 10-fold cross-validation. The setups where the difference is statistically significant in terms of performance (compared with the models trained with manual labels) are marked with an asterisk (*).

Noisy Labels	**F1 Labels**	**Model**	**Catania**	**RUMC**	**Cumulative**
Manual	-	CLAM_MB	0.617 ± 0.027	0.717 ± 0.023	0.674 ± 0.016
		transMIL	0.635 ± 0.024	0.745 ± 0.024	0.696 ± 0.016
		ViT	0.705 ± 0.033	0.812 ± 0.02	0.763 ± 0.012
1%	0.991	CLAM_MB	0.624 ± 0.022	0.725 ± 0.021	0.681 ± 0.014
		transMIL	0.634 ± 0.042	0.756 ± 0.012	0.700 ± 0.020
		ViT	0.697 ± 0.035	0.817 ± 0.018	0.762 ± 0.021
2%	0.98	CLAM_MB	0.621 ± 0.034	0.721 ± 0.016	0.677 ± 0.018
		transMIL	0.642 ± 0.033	0.739 ± 0.011	0.695 ± 0.017
		ViT	0.698 ± 0.032	0.807 ± 0.026	0.757 ± 0.026
5%	0.957	CLAM_MB	0.609 ± 0.035	0.715 ± 0.022	0.670 ± 0.021
		transMIL	0.622 ± 0.050	0.743 ± 0.015	0.687 ± 0.026
		ViT	0.699 ± 0.027	0.809 ± 0.029	0.758 ± 0.020
10%	0.907	CLAM_MB	0.601 ± 0.037	0.690 ± 0.034	0.653 ± 0.027
		transMIL	0.615 ± 0.029	0.739 ± 0.025	0.683 ± 0.023
		ViT	0.699 ± 0.026	0.808 ± 0.018	0.757 ± 0.015
20%	0.822	CLAM_MB	0.579 ± 0.060	0.725 ± 0.038	0.658 ± 0.042
		transMIL	0.614 ± 0.039	0.743 ± 0.017	0.684 ± 0.018
		ViT	0.702 ± 0.018	0.808 ± 0.015	0.759 ± 0.012
50%	0.561	CLAM_MB	0.409 ± 0.087*	0.528 ± 0.069*	0.477 ± 0.065*
		transMIL	0.483 ± 0.055*	0.566 ± 0.027*	0.537 ± 0.031*
		ViT	0.576 ± 0.049*	0.701 ± 0.040*	0.643 ± 0.038*

4.4. Colon cancer

The classification performance of multiple computer vision architectures trained with multilabel automatically annotated data to classify colon cancer WSIs is as effective as the performance reached by models using manually annotated data.

Tables 10 and 11 summarize the results. The highest performance using manual labels is reached using a ViT architecture (F1-score = 0.831 ± 0.009) on both test partitions, dramatically outperforming the other two ar-

Table 9: Results on the classification of lung cancer, in terms of F1-score. The performance is evaluated considering three computer vision architectures: CLAM, transMIL, ViT. The architectures are trained with automatic and manual weak multiclass labels, generated by extracting meaningful concepts from the corresponding pathology report using the SKET algorithm. The performance of SKET is reported in the 'Noisy Labels' column. The goal is to evaluate the effectiveness of automatic labels on the multiclass classification of WSIs. For every setup, the F1-score average and standard deviation of the classification performance are reported, considering the models trained with the 10-fold cross-validation. The setups where the difference is statistically significant in terms of performance (compared with the models trained with manual labels) are marked with an asterisk (*).

Noisy Labels	**F1 Labels**	**Model**	**Catania**	**RUMC**	**Cumulative**
Automatic	0.976	CLAM_MB	0.623 ± 0.031	0.705 ± 0.028	0.67 ± 0.020
		transMIL	0.620 ± 0.027	0.740 ± 0.027	0.686 ± 0.018
		ViT	0.682 ± 0.041	0.820 ± 0.014	0.756 ± 0.022
Manual	-	CLAM_SB	0.617 ± 0.027	0.717 ± 0.023	0.674 ± 0.016
		transMIL	0.635 ± 0.024	0.745 ± 0.024	0.696 ± 0.016
		ViT	0.705 ± 0.033	0.812 ± 0.020	0.763 ± 0.012

chitectures (CLAM reaches 0.773 ± 0.015, while transMIL reaches 0.791 ± 0.008).

Table 10 shows the classification performance obtained using multilabel manual labels and noisy labels. This experiment aims to investigate general rules for the adoption of automatic labels on the multilabel classification of WSIs. Considering all the architectures, the performance is similar to the one obtained using manual labels, especially until 20% of training samples are wrongly-annotated, the difference in terms of performance is not statistically significant. When the percentage of wrongly-annotated training is 50% the performance degrades and the difference, compared with manual labels, is statistically significant, suggesting this percentage of wrongly-annotated labels can be considered as a threshold for the adoption of automatic weak labels in a multilabel classification scenario.

Table 11 includes the comparison of automatic labels and manual labels. This comparison represents a real-case scenario of automatic data labeling, where automatic labels are generated by extracting concepts from reports. The comparison among labels shows a F1-score equal to 0.971, suggesting that the algorithm should lead to performance similar to the one obtained in the previous experiment using 2% and 5%. The results confirm the hyphotesis, since the performance are slightly worse than the one obtained using manual labels, but the gap is not statistically significant (according to

Table 10: Results on the classification of colon cancer, in terms of F1-score. The performance is evaluated considering three computer vision architectures: CLAM, transMIL, ViT. The architectures are trained with manual weak multilabel labels and with noisy weak labels, randomly perturbated according to different percentages of noise. The percentage of noisy labels is reported in the 'Noisy Labels' column, while the accuracy of the labels is reported in terms of F1-score, 'F1 labels' column. The goal is to evaluate the effect that noisy weak labels have on the multilabel classification of WSIs. For every setup, the F1-score average and standard deviation of the classification performance are reported, considering the models trained with the 10-fold cross-validation. The setups where the difference is statistically significant in terms of performance (compared with the models trained with manual labels) are marked with an asterisk (*).

Noisy Labels	**F1 Labels**	**Model**	**Catania**	**RUMC**	**Cumulative**
Manual	-	CLAM_MB	0.761 ± 0.015	0.780 ± 0.017	0.773 ± 0.015
		transMIL	0.771 ± 0.015	0.807 ± 0.007	0.791 ± 0.008
		ViT	0.824 ± 0.016	0.837 ± 0.007	0.831 ± 0.009
1%	0.988	CLAM_MB	0.761 ± 0.018	0.776 ± 0.016	0.771 ± 0.015
		transMIL	0.772 ± 0.014	0.810 ± 0.009	0.793 ± 0.010
		ViT	0.827 ± 0.018	0.835 ± 0.005	0.831 ± 0.009
2%	0.978	CLAM_MB	0.745 ± 0.018	0.764 ± 0.019	0.757 ± 0.017
		transMIL	0.777 ± 0.019	0.807 ± 0.010	0.793 ± 0.012
		ViT	0.821 ± 0.019	0.837 ± 0.005	0.831 ± 0.009
5%	0.943	CLAM_MB	0.765 ± 0.018	0.771 ± 0.021	0.769 ± 0.018
		transMIL	0.766 ± 0.013	0.808 ± 0.009	0.790 ± 0.008
		ViT	0.819 ± 0.015	0.835 ± 0.008	0.828 ± 0.009
10%	0.898	CLAM_MB	0.767 ± 0.023	0.777 ± 0.019	0.774 ± 0.018
		transMIL	0.768 ± 0.017	0.805 ± 0.009	0.789 ± 0.010
		ViT	0.827 ± 0.015	0.836 ± 0.005	0.833 ± 0.008
20%	0.814	CLAM_MB	0.748 ± 0.026	0.757 ± 0.020	0.754 ± 0.019
		transMIL	0.772 ± 0.012	0.809 ± 0.010	0.793 ± 0.008
		ViT	0.822 ± 0.020	0.833 ± 0.003	0.829 ± 0.009
50%	0.587	CLAM_MB	0.697 ± 0.042*	0.646 ± 0.086*	0.670 ± 0.056*
		transMIL	0.723 ± 0.027*	0.720 ± 0.024*	0.721 ± 0.015*
		ViT	0.811 ± 0.016*	0.804 ± 0.021*	0.807 ± 0.016*

Wilcoxon Rank-Sum test, comparing every setup to the one where manual labels are used).

5. Discussion

This paper evaluates the application of weak automatic labels to train computer algorithms on classification.

The application of automatic weak labels would dramatically reduce the

Table 11: Results on the classification of colon cancer, in terms of F1-score. The performance is evaluated considering three computer vision architectures: CLAM, transMIL, ViT. The architectures are trained with automatic and manual weak multilabel labels, generated by extracting meaningful concepts from the corresponding pathology report using the SKET algorithm. The performance of SKET is reported in the 'Noisy Labels' column. The goal is to evaluate the effectiveness of automatic labels on the multilabel classification of WSIs. For every setup, the classification performance's F1-score average and standard deviation are reported, considering the models trained with the 10-fold cross-validation. The setups where the difference is statistically significant in terms of performance (compared with the models trained with manual labels) are marked with an asterisk (*).

Noisy Labels	**F1 Labels**	**Model**	**Catania**	**RUMC**	**Cumulative**
Automatic	0.971	CLAM_MB	0.761 ± 0.014	0.771 ± 0.019	0.767 ± 0.016
		transMIL	0.759 ± 0.013	0.801 ± 0.004	0.783 ± 0.005
		ViT	0.813 ± 0.014	0.836 ± 0.008	0.826 ± 0.008
Manual	-	CLAM_MB	0.761 ± 0.015	0.780 ± 0.017	0.773 ± 0.015
		transMIL	0.771 ± 0.015	0.807 ± 0.007	0.791 ± 0.008
		ViT	0.824 ± 0.016	0.837 ± 0.007	0.831 ± 0.009

time needed to collect samples to train algorithms for the analysis of biomedical data. However, it is not clear under which conditions automatic labels can be adopted to train algorithms.

The results achieved in the paper show that automatic labels are as effective as manual ones, for the classification of WSIs. The first experiments (where manual labels are compared to different percentages of noisy labels) allow to identify some patterns in the algorithm performance. The noise introduced by mislabeled samples (inherently present within automatic labels) impacts the networks' accuracy and robustness. The performance of other samples may compensate for the effect of mislabeled samples on the training using the manual labels until a fixed percentage of mislabeled data: 10% regarding celiac disease (binary labels) and 20% regarding lung and colon cancer (respectively multiclass and multilabel labels). This performance decrease can be explained by considering the different natures of labels. Mislabeled samples have a high impact on binary classification since the label flipping leads to opposite results. Annotation errors are also disruptive in multiclass labels, even if, in this case, the effect can be smoothed if the errors involve similar classes (already prone to uncertainty). Another explanation for this gap can be identified in the training dataset size. Another relevant parameter to consider when automatic labels are applied is the size of the training dataset since the effect of mislabeled samples on the training may

be compensated by the other samples. In this paper, the celiac disease training dataset includes around 1'000 samples, while instead, the lung cancer dataset includes around 2'500 samples, and the colon cancer one includes around 6'500. In the celiac disease use case, when the percentage of mislabeled samples is 20% or more, the performance of the architectures is no longer comparable with the one reached using manual labels when the percentage of mislabeled samples is 20%. This result suggests that automatic labels can be adopted when the algorithm used to generate them is accurate. The effect of noisy labels can also be identified in the performance standard deviation: the higher the percentage of noisy labels, the less robust the three architectures are.

The architectures trained using automatic labels reach performance comparable (i.e. the performance difference is not statistically significant) with the one reached using manual labels. The results obtained using SKET to generate automatic weak labels show that automatic weak labels can be used to train different architectures on the classification of WSIs. The conditions identified using randomly perturbated noisy data are also tested on a real case scenario, where the automatic labels are generated using SKET, an NLP algorithm to extract meaningful concepts from pathology reports. This set of experiments is necessary to show the application of automatic labels in a real-case scenario, where the likelihood of mislabeling a sample varies. For example, if weak labels are automatically extracted from a report, depending on the report content, a sample has a higher chance of being mislabeled. This characteristic does not apply to the randomly perturbated noisy samples, where every sample can be randomly mislabeled.

The fact that automatic labels are as effective as manual labels opens many perspectives for the computational pathology domain and for the biomedical domain in general. Automatic labels limit the need for medical experts to annotate data, which can save up to 99.7% of time otherwise needed to analyze reports in order to infer labels. Therefore, a dataset that includes around 10,000 can be weakly-annotated in around five minutes. Considering that a large amount of biomedical data is produced every year and only a small percentage is annotated, this would allow the exploitation of a vast amount of data that can be used to build more accurate and robust models while still guaranteeing robust performance, helping medical experts diagnose diseases more effectively. The implementation details, such as architecture, task and pre-processing techniques adopted in this research, can be tailored to fit the specific characteristics of another problem.

6. Conclusions

The application of automatic labels may help exploit vast amounts of unlabeled biomedical samples to train more robust models, reducing by 99.7% the time needed to collect weakly annotated samples. However, it is still unclear when this label is effective. This paper evaluates the performance of different percentages of noisy labels (1,2,5,10,20,50%) and compares the results with the performance obtained by the same architectures but using manual weak labels provided by medical experts. After some rules are identified (e.g., training datasets with 10% of mislabeled samples lead to performance comparable to that obtained using manual labels), SKET, an algorithm for extracting meaningful concepts from reports, is used to generate automatic weak labels. The performance reached by the models trained with SKET labels is comparable (not a statistically significant difference) to the one obtained with manual labels, showing the effectiveness of automatic labels. The result can allow for the annotation of samples contained in hospitals without the need for human effort, paving the way for increasingly accurate algorithms. The code, including implementing the computer vision algorithms to classify WSIs, is publicly available on GitHub (https://github.com/ilmaro8/wsi_analysis).

Acknowledgments

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

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. The Journal of pathology 249, 286–294.

Benson, A.B., Venook, A.P., Al-Hawary, M.M., Cederquist, L., Chen, Y.J., Ciombor, K.K., Cohen, S., Cooper, H.S., Deming, D., Engstrom, P.F., et al., 2018. Nccn guidelines insights: colon cancer, version 2.2018. Journal of the National Comprehensive Cancer Network 16, 359–369.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: fast and flexible image augmentations. Information 11, 125.

Caio, G., Volta, U., Sapone, A., Leffler, D.A., De Giorgio, R., Catassi, C., Fasano, A., 2019. Celiac disease: a comprehensive current review. BMC medicine 17, 1–20.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, 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 medicine 25, 1301–1309.

Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G., 2018. Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognition 77, 329–353.

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, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.

Cifci, D., Veldhuizen, G.P., Foersch, S., Kather, J.N., 2023. Ai in computational pathology of cancer: improving diagnostic workflows and clinical outcomes? Annual Review of Cancer Biology 7, 57–71.

De Matos, J., Ataky, S.T.M., de Souza Britto Jr, A., Soares de Oliveira, L.E., Lameiras Koerich, A., 2021. Machine learning methods for histopathological image analysis: A review. Electronics 10, 562.

Deng, S., Zhang, X., Yan, W., Chang, E.I., Fan, Y., Lai, M., Xu, Y., et al., 2020. Deep learning in digital pathology image analysis: a survey. Frontiers of medicine 14, 470–487.

Fraggetta, F., Garozzo, S., Zannoni, G.F., Pantanowitz, L., Rossi, E.D., 2017. Routine digital pathology workflow: the catania experience. Journal of pathology informatics 8, 51.

Fraggetta, F., L'imperio, V., Ameisen, D., Carvalho, R., Leh, S., Kiehl, T.R., Serbanescu, M., Racoceanu, D., Della Mea, V., Polonia, A., et al., 2021. Best practice recommendations for the implementation of a digital pathology workflow in the anatomic pathology laboratory by the european society of digital and integrative pathology (esdip). Diagnostics 11, 2167.

Goode, A., Gilbert, B., Harkes, J., Jukic, D., Satyanarayanan, M., 2013. Openslide: A vendor-neutral software foundation for digital pathology. Journal of pathology informatics 4, 27.

Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B., 2009. Histopathological image analysis: A review. IEEE reviews in biomedical engineering 2, 147–171.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al., 2020. A survey on visual transformer. arXiv preprint arXiv:2012.12556 .

Hanna, M.G., Reuter, V.E., Ardon, O., Kim, D., Sirintrapun, S.J., Schüffler, 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. Modern Pathology 33, 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.

Hewer, E., 2020. The oncologist's guide to synoptic reporting: a primer. Oncology 98, 396–402.

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 clinical cancer informatics 3, 1–7.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, 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 .

Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2020. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical image analysis 65, 101759.

Krupinski, E.A., Graham, A.R., Weinstein, R.S., 2013. Characterizing the development of visual search expertise in pathology residents viewing whole slide images. Human pathology 44, 357–364.

Van der Laak, J., Litjens, G., Ciompi, F., 2021. Deep learning in histopathology: the path to the clinic. Nature medicine 27, 775–784.

Lebwohl, B., Rubio-Tapia, A., 2021. Epidemiology, presentation, and diagnosis of celiac disease. Gastroenterology 160, 63–75.

Litjens, G., Ciompi, F., van der Laak, J., 2022. A decade of gigascience: The challenges of gigapixel pathology images. GigaScience 11.

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. Nature biomedical engineering 5, 555–570.

Madabhushi, A., Lee, G., 2016. Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical image analysis 33, 170–175.

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. Journal of pathology informatics 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 digital medicine 5, 102.

Marini, N., Marchesin, S., Wodzinski, M., Caputo, A., Podareanu, D., Guevara, B.C., Boytcheva, S., Vatrano, S., Fraggetta, F., Ciompi, F., et al., 2024. Multimodal representations of biomedical knowledge from limited training whole slide images and reports using deep learning. Medical Image Analysis 97, 103303.

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. Frontiers in Computer Science 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. Journal of Pathology Informatics 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. Journal of Pathology Informatics 14, 100332.

Merchant, F., Castleman, K., 2022. Microscope image processing. Academic press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems 26.

Neumann, M., King, D., Beltagy, I., Ammar, W., 2019. Scispacy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv:1902.07669 .

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 .

Organization, W.H., 2023. Lung cancer. Online. URL: `https://www.who.int/news-room/fact-sheets/detail/lung-cancer`. accessed: 2024-04-25.

Schabath, M.B., Cote, M.L., 2019. Cancer progress and priorities: lung cancer. Cancer epidemiology, biomarkers & prevention 28, 1563–1579.

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. Advances in neural information processing systems 34, 2136–2147.

Sharir, G., Noy, A., Zelnik-Manor, L., 2021. An image is worth 16x16 words, what is a video worth? arXiv preprint arXiv:2103.13915 .

Travis, W.D., 2011. Pathology of lung cancer. Clinics in chest medicine 32, 669–692.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Wang, Y., Li, J., Metze, F., 2019. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 31–35.

Woolson, R.F., 2007. Wilcoxon signed-rank test. Wiley encyclopedia of clinical trials , 1–3.

Xu, H., Xu, Q., Cong, F., Kang, J., Han, C., Liu, Z., Madabhushi, A., Lu, C., 2023. Vision transformers for computational histopathology. IEEE Reviews in Biomedical Engineering .