Journal of Pathology Informatics 13 (2022) 100139

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

Original Research Article

Empowering digital pathology applications through explainable knowledge extraction tools

Stefano Marchesin [a], Fabio Giachelle [a], Niccolò Marini [b], Manfredo Atzori [b,c], Svetla Boytcheva [d], Genziana Buttafuoco [e], Francesco Ciompi [f], Giorgio Maria Di Nunzio [a], Filippo Fraggetta [e], Ornella Irrera [a], Henning Müller [b], Todor Primov [d], Simona Vatrano [e], Gianmaria Silvello [a,*]

[a] Department of Information Engineering, University of Padua, Padua, Italy
[b] Information Systems Institute, University of Applied Sciences Western Switzerland, Delémont, Switzerland
[c] Department of Neuroscience, University of Padua, Padua, Italy
[d] Sirma AI, Bulgaria
[e] Pathology Unit Gravina Hospital Caltagirone ASP Catania, Italy
[f] Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands

ARTICLE INFO

Keywords:
Clinical practice
Digital pathology
Expert systems
Explainable AI
Knowledge extraction
Machine learning

ABSTRACT

Exa-scale volumes of medical data have been produced for decades. In most cases, the diagnosis is reported in free text, encoding medical knowledge that is still largely unexploited. In order to allow decoding medical knowledge included in reports, we propose an unsupervised knowledge extraction system combining a rule-based expert system with pre-trained Machine Learning (ML) models, namely the Semantic Knowledge Extractor Tool (SKET). Combining rule-based techniques and pre-trained ML models provides high accuracy results for knowledge extraction. This work demonstrates the viability of unsupervised Natural Language Processing (NLP) techniques to extract critical information from cancer reports, opening opportunities such as data mining for knowledge extraction purposes, precision medicine applications, structured report creation, and multimodal learning.

SKET is a practical and unsupervised approach to extracting knowledge from pathology reports, which opens up unprecedented opportunities to exploit textual and multimodal medical information in clinical practice. We also propose SKET eXplained (SKET X), a web-based system providing visual explanations about the algorithmic decisions taken by SKET.

SKET X is designed/developed to support pathologists and domain experts in understanding SKET predictions, possibly driving further improvements to the system.

Introduction

Exascale volumes of multimodal data have been produced for decades in the biomedical domain. Biomedical data include patient information, clinical data, biological laboratory data, bio-images, bio-signals, instrumental examinations, and genetic data.

Hundred of thousands of reports have been used to describe findings leading to diagnoses, encoding vast medical knowledge. Free-text reporting is the standard for communicating the diagnosis, guiding patients' treatment, and other applications, such as cancer registries. Processing high volumes of free-text reports, usually performed manually, is also required to extract knowledge to train Machine Learning (ML) algorithms.

However, the manual analysis of data becomes an extremely time-consuming process since reports vary widely between institutions, might be written in languages other than English, contain noise, and do not present a standard structure. In this context, Natural Language Processing (NLP) methods are central[1–8] as they empower the efficient automatic processing of thousands of clinical reports and the extraction of key information for several downstream tasks, such as clinical note mining[9,10] and structuring,[11] risk prediction,[12] clinical decision-support,[13] and precision medicine retrieval.[14]

In the context of digital pathology, NLP techniques can drive noticeable advances by exploiting the availability of textual pathology reports paired with digital histopathology images (i.e., Whole Slide Images (WSIs)) in clinical practice. WSIs are used as a gold-standard to diagnose cancer cases and related diseases.[15,16] Within WSIs, tissue patterns and morphology vary depending on the image magnification level – enabling different tasks such as detection, classification, or segmentation.[17] However, the lack of training datasets containing pixel-wise annotations for entire images[18–20] limits the effectiveness of supervised ML models.[21]

* Corresponding author.
E-mail addresses: stefano.marchesin@unipd.it (S. Marchesin), gianmaria.silvello@unipd.it (G. Silvello).

http://dx.doi.org/10.1016/j.jpi.2022.100139
Received 16 August 2022; Received in revised form 6 September 2022; Accepted 7 September 2022
Available online 15 September 2022

Nevertheless, from the textual pathology reports, it is possible to extract key concepts (e.g., the diagnosis outcome) to annotate the associated WSIs. Although noisy, the extracted concepts can then serve as weak labels to train prediction models for image classification tasks.[22,23] However, even though automated solutions involving ML are increasingly being integrated into biomedical domains, NLP applications to digital pathology are less common. Compounding the situation further, the actual use of Artificial Intelligence (AI) algorithms in digital pathology requires a large amount of data annotations by pathologists. However, they are rarely available in a clinical setting.[1,24]

To overcome such limitation, this work aims at proving the viability of unsupervised NLP techniques to automatically extract critical information from pathology reports and use it for different digital pathology applications, such as automatic report annotation, pathological knowledge visualization, and WSI classification. In this regard, we present the Semantic Knowledge Extractor Tool (SKET), an unsupervised hybrid knowledge extraction system that combines an expert system with pre-trained ML models to extract knowledge from pathology reports. In recent years, NLP has shifted from using rules to ML approaches,[8,25] which have the advantage of learning regularities from data and of generalizing to previously unseen patterns.

Moreover, the advent of efficient Neural Language Models (NLMs)[26–29] paved the way for the pre-training era, where large NLMs trained in a self-supervised fashion on huge datasets are used to develop NLP models for a number of downstream tasks. Nevertheless, similarly to Santus et al.,[9] we argue that rule-based techniques capture critical information that should be used together with – and not substituted by – ML to improve performance.

We evaluate SKET effectiveness on entity linking and text classification, considering 3 different diseases: colon, cervix, and lung cancer. In this regard, we resort on diagnostic reports coming from 2 medical centers in Italy and The Netherlands. Then, we compare SKET with unsupervised ML approaches to understand the impact that combining rule-based techniques and pre-trained ML models have on the extraction of knowledge from pathology reports. The achieved results highlight the viability of ML methods for information extraction in the pathology domain, but also stress the importance of expert knowledge to reach the high levels of accuracy required to (semi-)automate the clinical practice. Moreover, the applicability of the proposed approach is enhanced by the considered multilingual setting.

Besides effectiveness, we must consider that understanding and explaining decisions and outcomes is crucial in clinical practice.

However, the black-box nature of many ML models, especially those based on Deep Learning, makes it difficult to understand and trace back the underlying decision process. Hence, there is an urgent need for a shift towards eXplainable Artificial Intelligence (XAI).[31,32,69,70]

In the biomedical domain, clinicians and domain experts need to understand *why* a specific output has been produced to trust the system and its predictions; moreover, the explainability of algorithmic decisions is increasingly required for legal reasons.[68] To this end, we propose SKET eXplained (SKET X). This web-based system allows domain experts to interact with SKET and visually comprehend the outcomes, rules, and parameters used in the knowledge extraction process. In addition, SKET X allows users to compare different SKET executions (e.g., with varying parameters of the system) and inspect punctual information, which provides valuable insights into the knowledge extraction process and the contribution of each component and parameter. SKET X aims to support pathologists and domain experts in the interaction with SKET, allowing them to gain an in-depth comprehension of the system decision process – and thus increasing trust and confidence in the system. Beyond explainability, we also report different digital pathology applications where SKET has been successfully integrated as a core system.[33,34] In particular, we deepen the use of SKET in such applications and the advantages it entails.

SKET source code is publicly available at https://github.com/ExaNLP/sket. Besides, SKET can also be deployed as a Docker container. For information about the Docker version of SKET, please refer to https://github.com/ExaNLP/sket#docker. SKET X is available at http://w3id.org/sketx[1].

The rest of this paper is organized as follows: Section 2 - "Material" describes the considered data resources. Section 3 - "Methods" presents SKET. Section 4 - "Evaluation" describes the experimental setup and reports quantitative and qualitative results. Section 5 - "Understanding through explainability" presents SKET X. Section 6 - "Digital pathology applications" outlines the digital pathology downstream applications empowered by SKET. Finally, Section 7 - "Conclusions and future work" draws some conclusions.

Material

The data used to develop and evaluate SKET comes from 2 different medical centers: the Cannizzaro Hospital (AOEC), Catania, Italy and the Radboud University Medical Center (RUMC), Nijmegen, The Netherlands. The AOEC data include diagnostic reports for colon, cervix, and lung cancer cases, written in Italian and associated with WSIs. All data were collected in the clinical workflow and fully anonymized afterwards. Similarly, the RUMC data consist of diagnostic reports and the associated WSIs for colon and cervix cases, written in Dutch – after the use of speech-to-text tools – and anonymized. For both medical centers, the considered reports have been provided directly in digital format and span several diagnostic outcomes. Note that other medical centers may provide reports in non-digital format, thus requiring a digitization step upstream of SKET. Table 1 reports the total number of diagnostic reports for each considered use-case and medical center.

Diagnostic reports contain the results of the analyses performed on specific tissues (or cells) to obtain a pathological–clinical diagnosis – i.e., presence or absence of the disease. AOEC and RUMC diagnostic reports follow the College of American Pathologists (CAP) international guidelines[2] for pathology reports[35,36] and contain the patient's personal and clinical-specific information, the description of how a specimen appears to the naked eye and at the microscope, and provide the final diagnosis.

As mentioned above, AOEC and RUMC diagnostic reports are written in Italian and Dutch, respectively. However, most of the resources required to develop NLP methods that extract concepts from unstructured text are in English. To overcome this limitation, we first translated diagnostic reports in English and then performed data curation over them. We used the open-source, pre-trained Marian Neural Machine Translation (NMT) models,[37] which exhibit a Transformer-based[38] encoder–decoder architecture with 6 layers in each component. Given the complexity of the task, such an automatic approach introduces systematic translation errors that, if propagated, could hamper the effectiveness of the extraction process. For this reason, we performed a data curation step, in which recurring, manually identified translation errors were corrected through the use of handcrafted rules.

We defined an ontology[3] for modeling the clinical reports in the digital pathology domain: ExaMode[4] ontology. Amongst other aspects not relevant for the current work, the ontology specifically defines the key concepts and

Table 1
Data size. For each medical center, we report the number of diagnostic reports associated with each use-case. The "–" symbol represents the lack of reports for a given use-case.

	Colon	Cervix	Lung
AOEC	1704	1777	1902
RUMC	2065	2350	–

[1] Access credentials for reviewing: demo/demo.
[2] https://www.cap.org/protocols-and-guidelines.
[3] https://w3id.org/examode/ontology/.
[4] ExaMode stands for "Extreme-scale Analytics via Multimodal Ontology Discovery & Enhancement" and is an H2020 project financed by the EU commission. More information can be found at: http://www.examode.eu/.

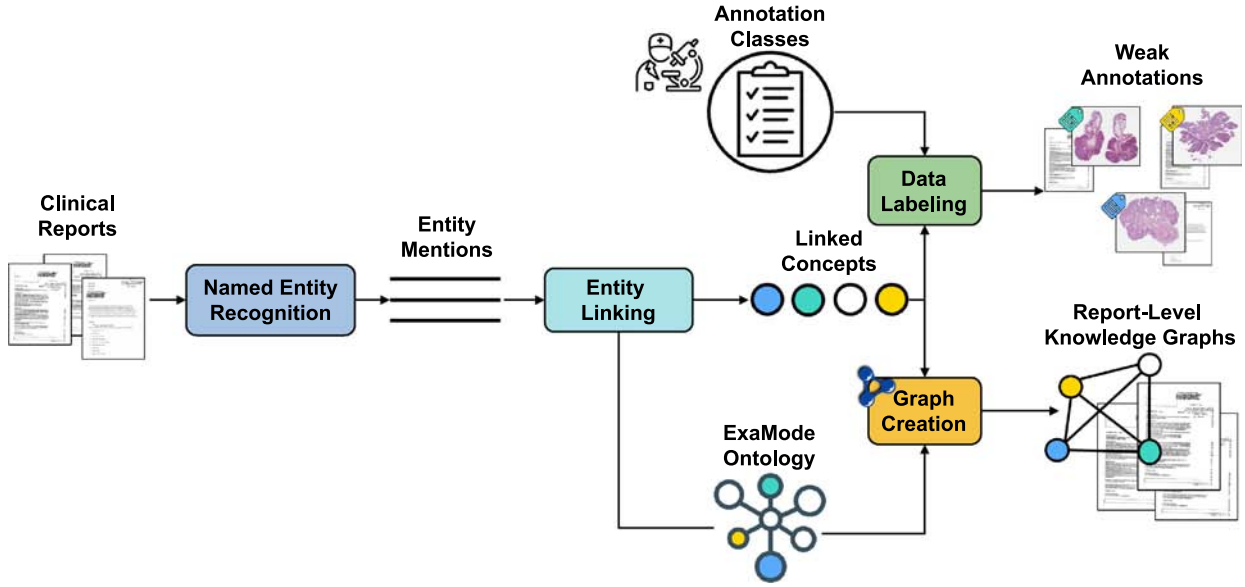


Fig. 1. SKET architecture. SKET main components are: (A) Named Entity Recognition, (B) Entity Linking, (C) Data Labeling, and (D) Graph Creation.

properties to model the diagnosis of colon, cervix, and lung cancer, the anatomical location where the disease might be located, the procedure employed to get the tissue, and the tests conducted on the tissue. Despite many medical ontologies focusing specifically on cancer exist, no single ontology comprehensively models all the diseases related to the cases mentioned above, their anatomical location, topography, and pathology laboratory process.

Methods

SKET adopts a combination of pre-trained Named Entity Recognition (NER) models and unsupervised Entity Linking (EL) methods to extract key concepts (entities) from the diagnostic reports and to link them to the reference ontology. The use of pre-trained NER models and unsupervised EL methods makes SKET suitable for weak supervision tasks. In this regard, the pathological concepts extracted from diagnostic reports can serve as weak labels to train prediction models for image classification tasks,[22,23] or as nodes to build report-level knowledge graphs for information retrieval tasks.[39]

As reported in Fig. 1, SKET consists of 4 components: (A) Named Entity Recognition, (B) Entity Linking, (C) Data Labeling, and (D) Graph Creation. Components (A) and (B) are sequential, whereas components (C) and (D) are parallel. Below, we describe for each component the different methods and techniques we adopted, expanded, or developed.

Named entity recognition

NER is the task of identifying and categorizing key information – i.e., entities – within text. An entity can be any word or phrase that consistently refers to the same concept or object of the world. Each identified entity is classified into a pre-defined category, such as disease, protein, gene, cell type, etc.

SKET relies on a combination of pre-trained neural models and rule-based techniques to perform NER. At its core, SKET adopts ScispaCy models,[40] which provide full NER pipelines for biomedical data, comprising large medical vocabularies, and Word2Vec[26] word vectors trained on the PubMed Central Open Access Subset.[41] It is worth mentioning that SKET has been designed to be deployed with any of the core models available at: https://allenai.github.io/scispacy/.

Then, SKET extends the ScispaCy pipeline with 2 additional components: Entity Fusion and Negation Detection.

Entity Fustion: SKET extends the NER pipeline with a set of rules used to identify and merge specific entities otherwise regarded as separate by ScispaCy. For instance, “transverse” and “colon” are considered as separate entities, whereas we are interested in “transverse colon” as a unique entity. Hence, we developed regular expressions that identify trigger terms indicative of a set of otherwise potentially separate entities. Once a trigger term is identified, SKET matches the entities extracted by ScispaCy with the candidate terms associated with the trigger. Depending on the trigger term, the match that SKET performs between extracted entities and candidate terms follows different rules based on directional and positional attributes. Directional attributes specify the set of extracted entities to be matched with the candidate trigger terms, and it can assume 3 values:

(1.) PRE: match with the entities preceding the trigger entity.
(2.) POST: match with the entities succeeding the trigger entity.
(3.) BOTH: match between the entities both preceding and succeeding the trigger entity.

Positional attributes specify the maximum distance allowed between the trigger entity and the other one, and it can assume 2 values:

(1.) EXACT: the matched entity must be right before/after the trigger entity.
(2.) LOOSE: the matched entity can be anywhere before/after the trigger entity.

The described set of rules has been developed on a holdout dataset and it is available on the SKET GitHub repository.[5] The dataset consists of 50 diagnostic reports for each use-case and medical center, for a total of 250 diagnostic reports.

Negation Detection: To handle negated entities, we extend the NER pipeline with NegEx,[42] a negation detection algorithm evaluating whether extracted entities are negated within text. NegEx uses regular expressions to identify the scope of trigger terms that are indicative of negation, such as “no” or “ruled out”. Then, the entities extracted within the scope of a trigger term are marked as negated. In this way, SKET identifies – and removes from the final list of extracted entities – those entities that NegEx regards as negated. For example, if we consider the phrase “free of dysplasia”,

[5] https://github.com/ExaNLP/sket/tree/main/sket/nerd/rules/.

NegEx identifies the trigger term “free of” and marks “dysplasia” as negated, which is then removed by SKET.

Entity linking

EL is the task of assigning unique meanings to entities mentioned within text. In other words, the objective of EL is to determine whether a given entity refers to a specific concept or object within a reference ontology.

SKET employs a combination of ad-hoc and similarity matching techniques to link the extracted entities to unique concepts within the ExaMode ontology. Given an extracted entity, SKET first tries to match it using ad-hoc matching and when it fails SKET employs similarity matching.

Ad-hoc matching: SKET uses regular expressions to identify trigger terms indicative of a specific ontological concept. Once a trigger term is identified, SKET matches the entity containing the trigger term with the closest ontology concept. For instance, if an extracted entity contains the term “carcinoma”, then SKET links the entity to the ontology concept “colon adenocarcinoma”. As for Entity Fusion, the ad-hoc matching rules have been developed on the holdout dataset and are available on GitHub.

Similarity matching: SKET performs similarity matching using a combination of string and semantic matching techniques. For string matching, SKET relies on the Gestalt Pattern Matching (GPM) algorithm,[43] which computes the similarity of 2 strings as the number of matching characters divided by the total number of characters in the 2 strings. Matching characters are those in the longest common subsequence plus, recursively, matching characters in the unmatched region on either side of the longest common subsequence. For semantic matching, SKET exploits the word vectors provided by ScispaCy models.[40] In other words, SKET performs semantic matching as the cosine distance between the vector representations of the extracted entities and the ontology concepts – where vector representations are the mean of the word vectors composing the extracted entities or the ontology concepts.

Both string and semantic matching produce a ranking of ontology concepts ordered by decreasing similarity with a given target entity. To combine the 2 rankings – and select the concept with the highest rank – SKET performs rank fusion using the CombSUM[44] with min–max normalization. Before selection, a pruning phase is performed on the combined ranking, in which ontology concepts with a similarity score lower than a pre-determined threshold are removed. The threshold value has been set empirically to 1.8 using the holdout dataset. The pruning phase aims to increase precision by reducing false positives, which occur when ontology concepts are incorrectly linked to the extracted entities.

Data labeling

SKET also provides labels as one of its main outputs. Given the set of concepts extracted from each diagnostic report, SKET maps a clinically relevant subset of such concepts to a set of annotation classes defined by AOEC pathologists. For each use-case, we report below the set of annotation classes.

Colon annotations: (1) Cancer; (2) Adenomatous polyp - high grade dysplasia; (3) Adenomatous polyp - low grade dysplasia; (4) Hyperplastic polyp; (5) Non-informative.

Cervix annotations: (1) Cancer - adenocarcinoma in situ; (2) Cancer - adenocarcinoma invasive; (3) Cancer - squamous cell carcinoma in situ; (4) Cancer - squamous cell carcinoma invasive; (5) High grade dysplasia; (6) Low grade dysplasia; (7) HPV infection present; (8) Koilocytes; (9) Normal squamous; (10) Normal glands.

Lung annotations: (1) Cancer - non-small cell cancer, adenocarcinoma; (2) Cancer - non-small cell cancer, large cell carcinoma; (3) Cancer - non-small cell cancer, squamous cell carcinoma; (4) Cancer - small cell cancer; (5) No cancer.

Thus, the Data Labeling component produces annotations from diagnostic reports that can be used to perform weakly supervised classification tasks.

Graph creation

SKET also builds report-level knowledge graphs using the extracted concepts as nodes and the semantic relations of the ExaMode ontology as edges. The use of ontology concepts and relations to describe diagnostic reports increases the semantic understanding of the underlying data.[45] Once created, report-level knowledge graphs are encoded in a machine-readable format through Resource Description Framework (RDF).

Evaluation

Tasks

We evaluate the effectiveness of SKET on Entity Linking (Task 1) and Text Classification (Task 2). The evaluation of SKET on entity linking also serves as a proxy to validate the quality of the RDF graphs it produces. On the other hand, text classification results help understanding the viability of using SKET as an automatic annotator in weak supervision tasks. Between the 2 tasks, text classification has a prominent role as it provides weak annotations that can be used to reduce the high costs of training cancer assisted diagnosis tools – which prevent unleashing the full potential of digital pathology applications.[34]

Datasets

Entity linking (Task 1): We evaluate SKET effectiveness to extract concepts from pathology reports on a subset of the proprietary data described in Section 2 - "Material". For each use-case and medical center, 250 reports have been manually annotated by experts using the concepts from the ExaMode ontology. Overall, the total number of annotated reports amounts to 1250. In terms of annotations, all use-cases have been annotated with a large number of different concepts. For colon cancer, the number of different concepts that can be found within reports stands at 19, while for cervix and lung cancer amounts to 21 and 11, respectively. This large number of different concepts highlights the complexity of the task, both for model predictions and human annotation efforts. In particular, the task can be seen as an extreme multi-label classification problem,[46,47] where the goal is to tag a given report with a subset of the relevant concepts from a large concept list.

Text classification (Task 2): To evaluate the effectiveness of SKET to weakly annotate pathology reports, the proprietary data described in Section 2 - "Material" has been manually labeled by experts using the annotation classes defined by AOEC pathologists. For each use-case, AOEC and RUMC reports have been annotated with one or more classes, making the task a multi-label classification problem. Table 2 reports the total number of reports annotated for each class in each use-case. Given the multi-label nature of the task, the total number of annotations does not reflect the total number of reports. As a side note, the class imbalance of the datasets reflects a real-case scenario, where certain conditions – e.g., low-grade dysplasia in colon cases – occur more often than others in the clinical routine.

Baselines

Entity linking (Task 1): We compare SKET with 2 unsupervised approaches based on Bio FastText[27,48] and BioClinical BERT[29,49] models. For a fair comparison, both approaches adopt the same NER ScispaCy pipeline used by SKET, but without the extensions introduced with it. Then, the approaches perform EL by computing the cosine distance between the vector representations of the extracted entities and the ontology concepts – obtained with FastText in one case and with BERT in the other. The ontology concept closest to the extracted entity is kept and, when appropriate, mapped to the corresponding annotation class. Both methods represent a straightforward approach to perform text classification with lack of annotated data.

Text classification (Task 2): We compare SKET with the Bio FastText and BioClinical BERT unsupervised approaches described above. Beyond unsupervised approaches, we also use SKET to weakly annotate diagnostic

Table 2
Number of annotated diagnostic reports for each use-case. Label counts are independent of each other except for "Non-informative" in colon, "Normal squamous" and "Normal glands" in cervix, and "No cancer" in lung, which only occur when none of the others does.

Colon	
Cancer	495
Adenomatous polyp – high-grade dysplasia	510
Adenomatous polyp – low-grade dysplasia	841
Hyperplastic polyp	508
Non-informative	1140
Cervix	
Cancer - adenocarcinoma in situ	125
Cancer - adenocarcinoma invasive	32
Cancer - squamous cell carcinoma in situ	638
Cancer - squamous cell carcinoma invasive	88
High-grade dysplasia	1544
Low-grade dysplasia	1053
HPV infection present	1221
Koilocytes	86
Normal squamous	1265
Normal glands	1266
Lung	
Cancer - non-small cell cancer, adenocarcinoma	961
Cancer - non-small cell cancer, large cell carcinoma	68
Cancer - non-small cell cancer, squamous cell carcinoma	528
Cancer - small cell cancer	144
No cancer	247

reports and then train FastText and BERT models in a supervised fashion. In this case, we stack a classification layer on top of the pre-trained models and perform end-to-end classification – i.e., the models take diagnostic reports as input and directly produce classes as output. Due to the introduction of supervised models, performances on text classification are obtained through 10-fold cross-validation.

Results

Entity linking (Task 1): Table 3 reports the results obtained by SKET and the considered baselines on entity linking. Overall, we see that SKET achieves high performances for both micro- and weighted-average F1 measures in each use-case. As for accuracy, the performances vary depending on the use-case, and the lowest score is obtained in colon cancer with a value of 0.6280. In terms of use-cases, the best SKET results are obtained on lung cancer. Compared to colon and cervix cases, lung cancer presents a lower number of concepts to identify, thus reducing the task complexity. On the other hand, colon and cervix use-cases show similar SKET performances, having a comparable number of concepts.

Table 3
Entity linking results on colon, cervix, and lung cancer pathology reports. The considered measures are subset accuracy, micro F1, and weighted F1. **Bold** values represent the highest scores achieved for each measure.

Approach	Model	Measures		
		Accuracy	Micro F1	Weighted F1
Colon				
Unsupervised	SKET	**0.6280**	**0.8861**	**0.8694**
	FastText	0.0660	0.5000	0.6146
	BERT	0.1840	0.3905	0.4527
Cervix				
Unsupervised	SKET	**0.7020**	**0.8322**	**0.8368**
	FastText	0.0900	0.2802	0.3439
	BERT	0.0720	0.2715	0.2940
Lung				
Unsupervised	SKET	**0.8624**	**0.9375**	**0.9262**
	FastText	0.2510	0.5610	0.6506
	BERT	0.3806	0.6804	0.8395

Table 4
Text classification results on colon, cervix, and lung cancer pathology reports. The considered measures are subset accuracy, micro F1, and weighted F1. The † symbol represents the statistical difference of SKET from unsupervised FastText- and BERT-based approaches – verified using a paired t-test with a p-value < 0.01. **Bold** values represent the highest scores achieved for each measure.

Approach	Model	Measures		
		Accuracy	Micro F1	Weighted F1
Colon				
Unsupervised	SKET	0.7525†	0.8386†	0.8373†
	FastText	0.4146	0.5298	0.5514
	BERT	0.5167	0.5697	0.6587
Weakly supervised	FastText	0.7116	0.8287	0.8276
	BERT	**0.7586**	**0.8432**	**0.8421**
Cervix				
Unsupervised	SKET	0.5281†	0.7791†	0.7611†
	FastText	0.2533	0.4882	0.4445
	BERT	0.3066	0.3962	0.4867
Weakly supervised	FastText	0.4744	0.7542	0.7566
	BERT	**0.5397**	**0.7901**	**0.7737**
Lung				
Unsupervised	SKET	0.8137	0.8387	0.8262
	FastText	0.5221	0.7296	0.6853
	BERT	**0.8523**†	**0.8630**†	**0.8526**†
Weakly supervised	FastText	0.7701	0.8313	0.8247
	BERT	0.8127	0.8375	0.8249

When we compare SKET performances with unsupervised approaches, we can see that SKET outperforms them for all measures in each use-case. This result shows the effectiveness of combining ad-hoc rules with ML models, which make SKET both precise and sensitive. Indeed, ad-hoc matching makes SKET precise while semantic matching makes it sensitive. To further support this outcome, we observe that the performances of unsupervised baselines – only relying on ML models and semantic matching – have low accuracy values. Given that we consider entity linking as a multi-label task, we resort on subset accuracy – where the set of concepts predicted for a report must exactly match the corresponding set of ground-truth concepts. Thus, accuracy values are more prone to rapidly decreasing with a large number of classes, and less precise models are naturally affected by this behavior.

Text classification (Task 2): Table 4 reports the results obtained by SKET and the considered baselines on text classification. Overall, we observe that SKET achieves high performance on colon and lung cancer use-cases, whereas it shows low accuracy values on cervix cancer. The motivation behind this drop in performance on cervix reports can be attributed to the high number of annotation classes (i.e., 10) and the multi-label setting. We recall that we rely on subset accuracy, which performs exact match between predicted and ground-truth labels – causing performance to drop faster when the number of classes is larger. The higher values for both micro- and weighted F1 measures, which do not perform exact match between predicted and ground-truth labels, further support this intuition.

Compared to unsupervised baselines, SKET achieves better performance in both colon and cervix use-cases. In particular, the (relative) performance gap between SKET and baselines varies from 20% to 40% across measures. To confirm SKET effectiveness, we conducted a paired t-test and found that there is a statistical difference (p-value < 0.01) between its performance and that of the baselines on all the considered measures.

This outcome shows the effectiveness of introducing ad-hoc rules at both NER and EL levels, as well as the soundness of combining different matching techniques together. On the other hand, the unsupervised BERT-based approach outperforms both SKET and FastText in lung cancer. In this case, the paired t-test confirmed a statistical difference between BERT performance and that of SKET and FastText. Nevertheless, the performance gap between BERT and SKET never exceeds 5%. This highlights the robustness of SKET across different use-cases and makes it a viable solution

in real scenarios, where annotated data are hard and expensive to get (such as in clinical practice). Besides, the lung cancer use-case presents 2 major differences with colon and cervix ones. First of all, lung annotation classes all revolve around different, but closely related, cancer types. As a consequence, contextualized NLMs (e.g., BERT[29]) – which are able to properly model the small semantic, contextual variations of such classes – achieve competitive results. Secondly, lung cancer data only consists of AOEC reports. The lack of RUMC reports makes the dataset more homogeneous and easier than the others, thus reducing classification inconsistencies for baseline models too.

Regarding weakly supervised models, the results reported in Table 4 demonstrate the effectiveness of using SKET to weakly annotate diagnostic reports and then train FastText and BERT models in a supervised fashion. In this regard, both weakly supervised FastText- and BERT-based approaches outperform their unsupervised counterparts. The only exception is for BERT on lung cancer data, where the unsupervised BERT approach achieves top performance. On the other hand, the weakly supervised BERT obtains the best results overall in both colon and cervix use-cases. Hence, SKET proves to be effective when used to bootstrap supervised models in absence of manual annotations. Following this procedure, supervised models can first be trained on data automatically annotated by SKET and then fine-tuned on small manually annotated batches, thus reducing annotation times and costs.

Understanding through explainability

In recent years, the application of AI algorithms in the biomedical domain has experienced unprecedented growth[30,50,51] – especially to perform clinical decision-support and diagnostic activities.[52–54] Therefore, there is an urgent need for eXplainable Artificial Intelleigence (XAI) tools that can help clinicians and domain experts understand algorithm predictions and their underlying rationale. In this regard, explainability techniques highlight decision-relevant aspects of algorithms that contribute to specific predictions, thus trying to answer why a model has made a certain decision.[31,55,56] Hence, explainability methods are essential for humans – and in particular for clinicians – to decide whether to trust algorithm predictions and the (underlying) models that generated them.[69,70] Among its different uses, explainability can be employed to understand the rationale of NER and EL outputs – such as the entity mentions and concepts identified by SKET within clinical reports.

However, since most of the data that humans can easily "visualize" regards objects restricted to the two/three-dimensional space, there is an urgent need not only for explainable models but also for explanation interfaces.[31] To this end, we have developed SKET X,[6] a web-based environment to interact with SKET and get useful insights about the extraction process and the related outputs. Through SKET X, pathologists and domain experts can visually comprehend SKET and the different components activated during the knowledge extraction process – thus getting a point-wise explanation of the outputs obtained for the provided diagnostic reports.

SKET X exploits Visual Analytics (VA) techniques to support domain experts in the visual comprehension of SKET outputs by means of intuitive and interactive interfaces. Such interfaces allow users to inspect and find out non-evident patterns in data and take decisions accordingly.[57] Specifically, VA techniques enable users to visually comprehend the results of an ongoing task, while it advances asynchronously in the background. Because of this, VA techniques are also used to visually adjust the parameters of a model running as a background task to continuously refine its outputs.[58,59]

SKET X workflow

SKET X is an interactive Webapp that runs SKET on a set of uploaded reports. SKET X is based on SKET pipelines definable by the user who can customize the parameters and run SKET multiple times to compare the outputs and all the intermediate steps of the process. Each pipeline runs as an asynchronous task, handled by a scheduler with a queue manager. The pipelines are organized for straightforward access in the dashboard interface, shown in Fig. 2. The dashboard provides information about the SKET pipelines executed by the users and enables access and download of the SKET outputs.

The execution of a SKET X pipeline consists at most of 3 *phases*, where the currently selected stage is shown on the top of the interface (see Fig. 3.A).

- *Translation*: The reports are automatically translated from their original language to English. Fig. 3 reports the information contained in the *Overview* tab of the interface, i.e. the inputs, outputs, and parameters of the translation phase.
- *Entity linking*: The entities automatically recognized within the reports are linked to the concepts in the ExaMode ontology. This phase's output consists of the identified mentions and the linked concepts. SKET employs a combination of hand-crafted rules and pre-trained neural models in this phase. The rules relevant to the disease of the given report are shown via a Sankey diagram, where the rules activated for the current report are highlighted. In this context, a rule is activated when one of the identified mentions – e.g., low degree dysplasia (mild) – satisfies one rule trigger – e.g., dysplasia && mild – that implies a link to a specific concept – e.g., mild colon dysplasia – as shown in Fig. 4.
- *Classification*: SKET exploits mapping rules to decide the appropriate labels for each report. As for the EL phase, the rules relevant for the disease of the considered report are visualized using a Sankey diagram, where the activated rules are highlighted. A rule is activated when one of the identified concepts – e.g., Mild Colon Dysplasia – satisfies one rule trigger – e.g., dysplasia && mild – that implies a specific label – e.g., Adenomatous polyp - low grade dysplasia – as shown in Fig. 5. The mentions and concepts considered in the classification task are regarded as *key* mentions/concepts, whereas the ones not satisfying any rule trigger are regarded as *excluded*, as shown in Fig. 5.C and 5.D, respectively. For instance, in Fig. 5 we can observe that the key concepts identified are Colon Hyperplastic Polyp and Mild Colon Dysplasia, whereas the excluded ones are Biopsy of Colon and Colon, NOS – both related to the same excluded mention colon biopsy.

SKET X interface

The interface of SKET X consists of 6 tabs providing different views of the data, according to the selected phase:

- *Overview tab*: Overview of the visual outputs available in the other tabs (i.e., *Input*, *Output*, *Params*, *Analytics*) for the current phase. The contents of the *Analytics* tab are shown in the overview only for the EL and classification phases.
- *Input tab*: It reports the input data for the current phase. For instance, if the considered phase is translation, this tab shows the reports in the original language, as shown in Fig. 3.B. Instead, if the considered phase is EL, it shows the translated reports. Similarly, the mentions and the concepts extracted for each report are shown in this tab for the classification phase.
- *Output tab*: It reports the output data for the current phase. For translation, this tab shows the reports translated into English, as shown in Fig. 3.C. Instead, if the considered phase is EL, it shows the mentions and the concepts extracted for each report. Similarly, the labels generated for each report are shown in this tab for the classification phase.
- *Params tab*: It reports the parameters for the current phase, as shown in Fig. 3.D. For instance, for EL, it shows the methods and models used by SKET to perform the linking process between mentions and related concepts. Another important parameter is the *threshold* used by SKET in the pruning phase to reduce false positives and thus increase precision, as described in Section 3.2 - "Methods, Entity Linking". When the phase considered is EL, users can change one or more parameters and then re-run SKET. This is useful to compare 2 pipelines using different parameters.

[6] http://w3id.org/sketx access provided with username and password: demo.

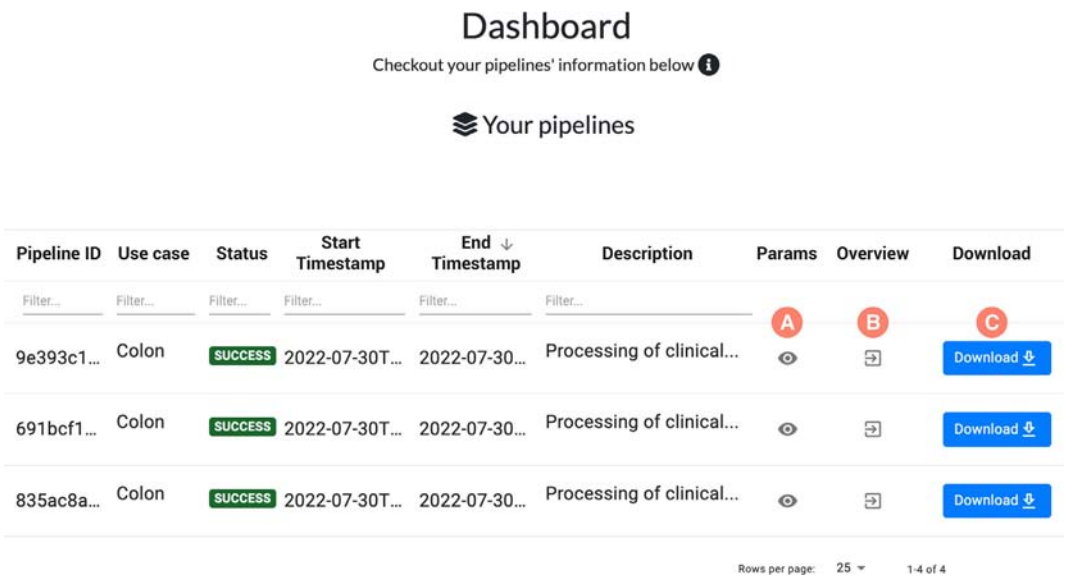


Fig. 2. SKET X dashboard providing information about the executed SKET pipelines - i.e., pipeline id, use case, pipeline status, start timestamp, end timestamp, description, pipeline parameters. Users can view the parameters of each pipeline by clicking on the dedicated button (A). Similarly, users can access pipelines data by clicking on the dedicated button (B). When the execution of a pipeline ends, its outputs become available for download (C).

- *Analytics tab*: It allows the users to analyze the current report's mentions, concepts, and labels in detail. In particular, if the considered phase is EL, users can inspect the identified mentions and concepts concerning the report textual content, as shown in Fig. 4.A and C. Moreover, by clicking on a mention, the user can inspect the list of associated concepts. At the same time, a user can also do the reverse - identifying the relevant mentions for a given concept. In addition, if the considered phase is classification, this tab shows the labels determined by SKET and the relations between a label and the concepts from which it derives, as shown in Fig. 5. To visually explain the rules used by SKET to determine both the concepts and labels, a Sankey diagram is reported on the right side of the interface as depicted in Figs. 4.B and 5.B. On the left side of the Sankey diagram,

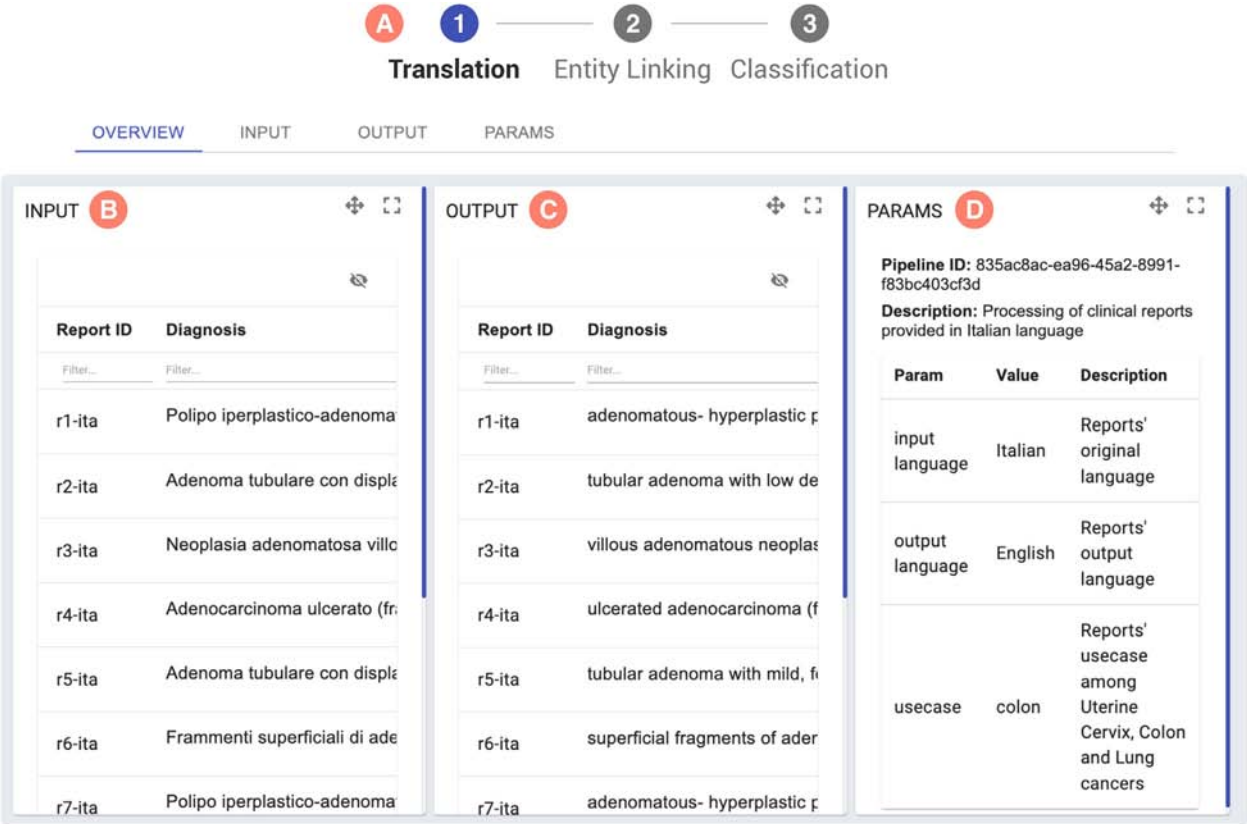


Fig. 3. (A) SKET X *Overview* tab for the translation phase, (B) the reports in the original language (input), (C) the translated reports (output) (C), and (D) the parameters and settings for the current phase.

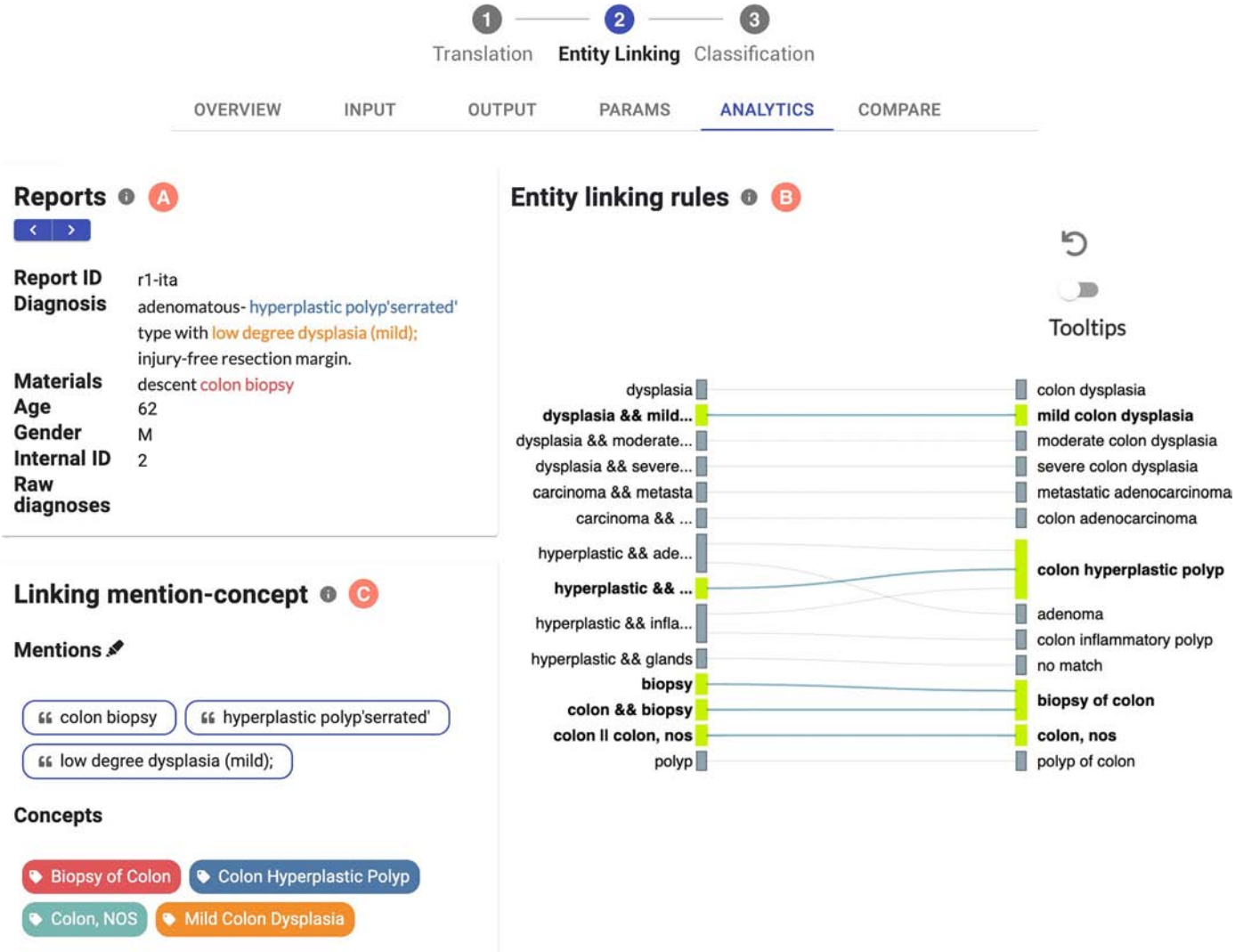


Fig. 4. SKET X *Analytics* tab for the EL phase: (A) reports section, the users can change the current report using the left/right buttons; (B) SKET rules for the NER task; and, (C) list of mentions and concepts produced by the knowledge extraction process. Each concept and related mentions are highlighted with the same color in (A) and (C). By clicking/hovering on a specific concept, it is possible to highlight the relevant rules in the Sankey diagram that determined the concept and the related mentions in the report text. On the left side of the Sankey diagram are reported the rules *triggers*, which are boolean expressions tested on each mentioned text. If one or more mentions satisfy a rule trigger, then the related concepts on the right side of the Sankey diagram are highlighted and listed in (C).

the rules *triggers* are reported, which are boolean expressions tested on the text of each mention – for the EL phase – and concept – for the classification phase. If one or more mentions/concepts satisfy a rule trigger, then the related concepts/labels on the right side of the Sankey diagram are highlighted.

- *Compare tab*: It allows the users to compare the outputs of 2 different SKET X pipelines in terms of mentions, concepts, and labels identified for the current report. When the users click on the *compare* tab, they are provided with an initial menu that allows them to specify the 2 pipelines to compare. After the selection, users can click on the *compare* button to visualize the interface dedicated for the comparison, illustrated in Fig. 6. The comparison interface is divided into 4 parts: (A) the reports section displaying information about the current report and 2 buttons for switching to the next/previous report; (B) the parameters section displaying pipeline information, such as the identifier, the description, and its parameters; (C) first pipeline section showing the outputs for the phase selected (e.g., mentions and concepts) and (D) second pipeline section with the same structure of (C). In particular, if the considered phase is EL, users can compare the concepts and the mentions identified by each pipeline and deduce which parameters have determined the major differences (e.g., the threshold for the NER task). Moreover, by clicking/hovering on each mention, the users can inspect the list of associated concepts (highlighted in different colors) among the 2 pipelines. On top of that, the common mentions between the 2 considered pipelines can be highlighted, thus making them and their related concepts easy to identify. Fig. 6 shows the outputs of 2 SKET pipelines that have been executed with different models, where the first pipeline uses only the neural model while the second one uses only GPM. Since the 2 pipelines considered in Fig. 6 use different models, they identify different concepts and mentions. Indeed, the common concepts between the 2 pipelines – i.e., Biopsy of Colon, Colon Hyperplastic Polyp, Colon NOS, and Mild Colon Dysplasia – have been identified using SKET rules, which are used in both pipelines. On the other hand, the disjoint concepts have been identified using the neural model – for Rectal mucous membrane – and GPM – for Adenoma and Resection – respectively. If the considered phase is classification, users can also compare the labels generated by each pipeline and the *key* concepts associated – i.e., the ones from which the labels are determined. For instance, in Fig. 7, we can observe that the labels generated by SKET are Adenomatous polyp – high grade dysplasia and Hyperplastic polyp for the first pipeline (C) while only Adenomatous polyp - high grade dysplasia for the second one (D). By clicking/hovering on the

1 Translation — 2 Entity Linking — 3 **Classification**

OVERVIEW | INPUT | OUTPUT | PARAMS | ANALYTICS | COMPARE

Reports A

Report ID	r1-ita
Diagnosis	adenomatous- hyperplastic polyp'serrated' type with low degree dysplasia (mild); injury-free resection margin.
Materials	descent colon biopsy
Age	62
Gender	M
Internal ID	2
Raw diagnoses	

Classification C

Labels

Adenomatous polyp - low grade dysplasia

Hyperplastic polyp

Key Mentions

“ hyperplastic polyp'serrated'

“ low degree dysplasia (mild);

Key Concepts

Colon Hyperplastic Polyp | Mild Colon Dysplasia

Excluded Mentions D

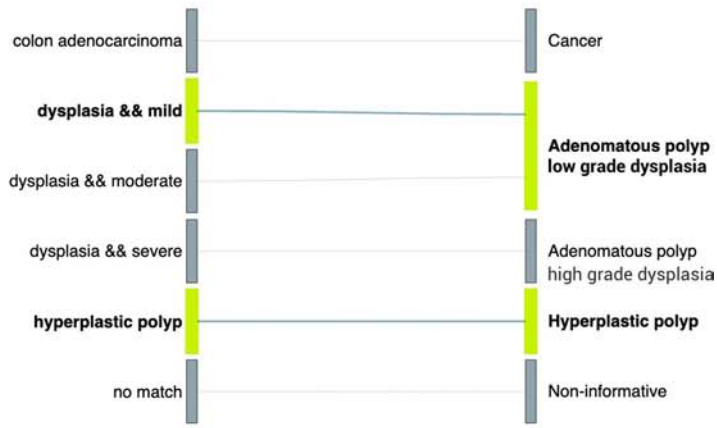
“ colon biopsy

Excluded Concepts

Biopsy of Colon | Colon, NOS

Concepts-labels association rules B

Tooltips



(caption on next page)

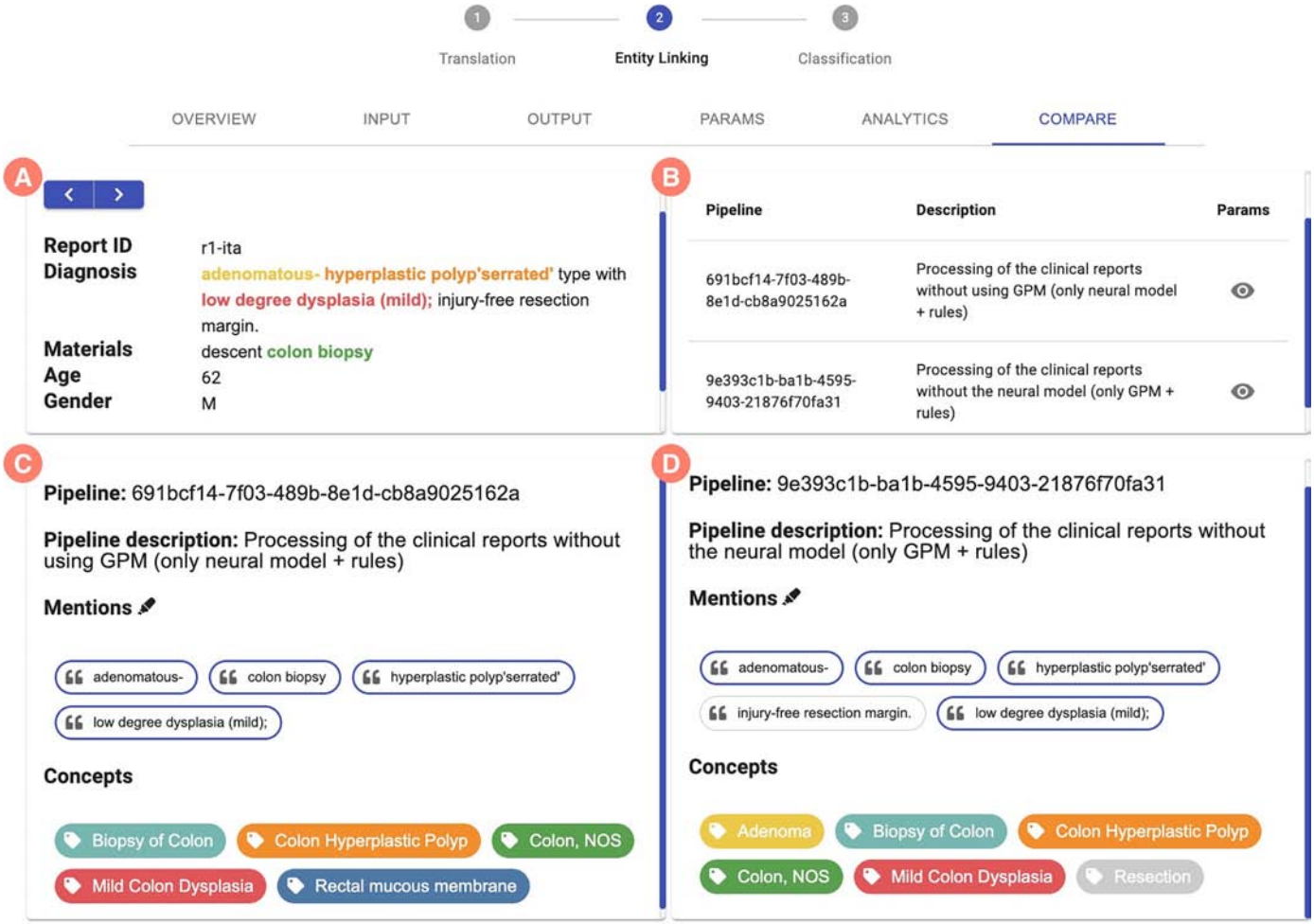


Fig. 6. SKET X *Compare* tab for the EL phase showing the comparison interface for the 2 pipelines specified for the comparison. The interface is organized in 4 parts: (A) the reports section displaying information about the current report and 2 buttons for switching to the next/previous report; (B) the parameters section displaying pipeline information, such as the identifier and its description, and parameters (e.g., the models used for EL phase and the threshold); (C) first pipeline outputs for the phase selected (e.g., mentions and concepts); and (D) second pipeline outputs for the phase selected. The mentions in common, and the related concepts, are highlighted both in the report text (A) and also in the mention/concept lists for each pipeline (C) and (D). Hence, we can observe that there is a mention injury-free resection margin and a concept Resection that are not highlighted since they have been identified only by the second pipeline (D). Nevertheless, the concepts Rectal mucous membrane and Adenoma have been identified only by respectively the first pipeline (C) and the second one (D), but since both are associated with the same common mention – i.e., adenomatous – they are highlighted as well.

Hyperplastic polyp label, users can realize that it derives from the Colon Hyperplastic Polyp concept, which, in turn, is associated with the polyp sigmoid mention. Nevertheless, the latter mention does not suggest the presence of a Colon Hyperplastic Polyp. Thus, it is a false positive. Similarly, users can do the same with the Adenomatous polyp - high-grade dysplasia label, discovering that it derives from the Severe Colon Dysplasia concept, which is correctly associated, through a SKET rule, with the severe dysplasia mention. Finally, users can also compare the *excluded* mentions and concepts that are not considered for the label generation process, but that can be a good indicator to determine whether the chosen threshold for models produces noisy concepts.

Hence, using SKET X pathologists and domain experts can visually comprehend why a certain concept/label has been extracted. Moreover, by leveraging both inspection and comparison functionalities, users can also understand the impact of different parameters on the obtained outputs, and thus investigate the advantages of combining ad-hoc rules with ML models to improve the overall effectiveness of knowledge extraction systems.

Digital pathology applications

SKET has been integrated as a core system into different downstream applications for digital pathology. Fig. 8 depicts the SKET ecosystem,

← **Fig. 5.** SKET X *Analytics* tab for the classification phase: (A) reports section to select the current report via left/right buttons; (B) SKET rules for determining the labels visualized with a Sankey diagram; and, (C) list of labels, mentions, and concepts determined by SKET. Each concept and the related mentions are highlighted with the same color in (A) and (C). The Sankey diagram highlights the relevant rules by clicking/hovering on a specific label. On the left side of the Sankey diagram are reported the rules *triggers*. If one or more concepts satisfy a rule trigger, then the related label is highlighted on the right side of the Sankey diagram and also listed in (C). The mentions and concepts involved in the classification task are the *key* mentions/concepts (C), while the *excluded* ones are reported in (D).

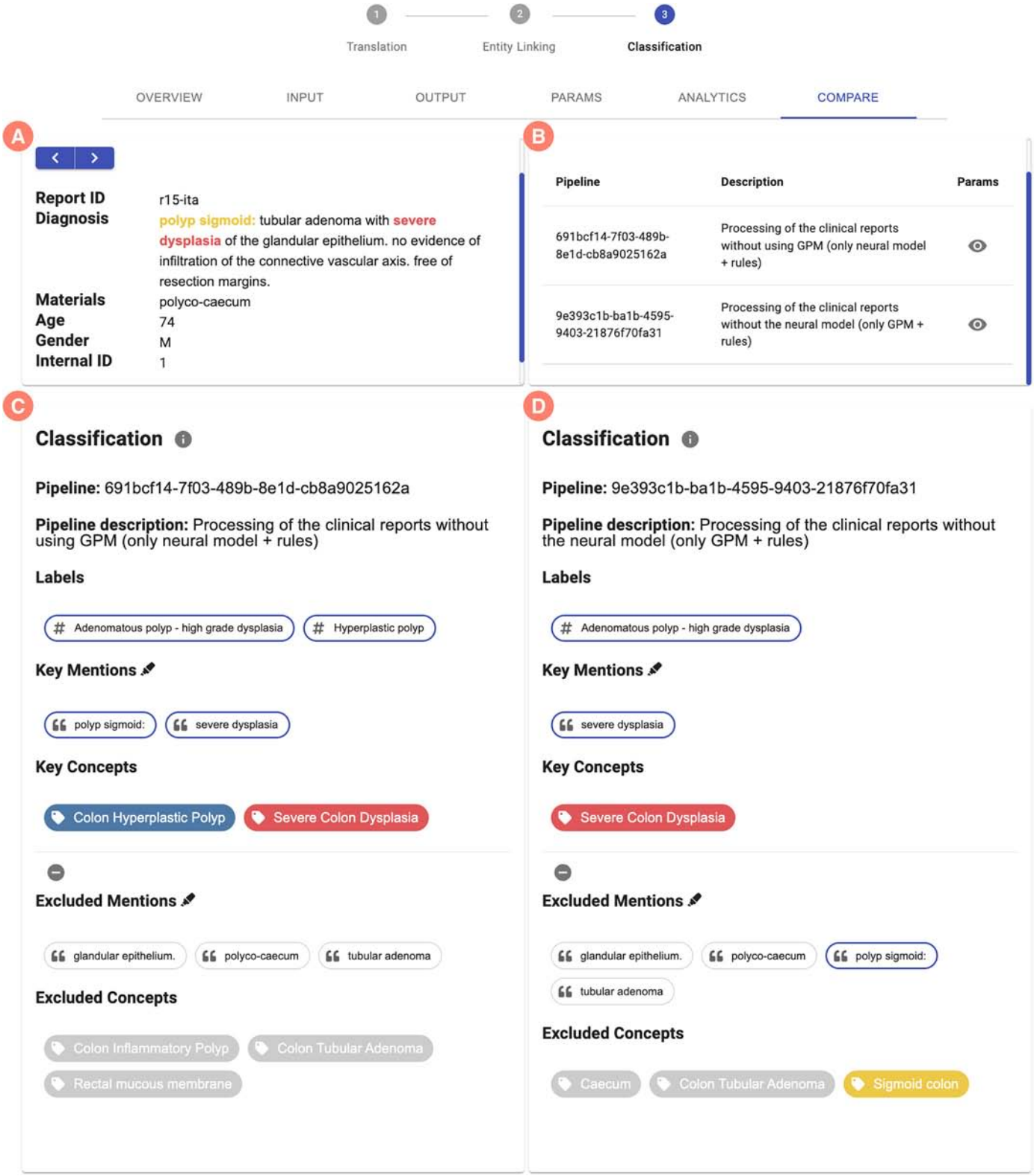


Fig. 7. SKET X *Compare* tab for the classification phase showing the comparison interface for the 2 pipelines specified for the comparison. The interface is organized in 4 parts: (A) the reports section displaying information about the current report and 2 buttons for switching to the next/previous report; (B) the parameters section displaying pipeline information, such as the identifier, the description, and its parameters; (C) first pipeline outputs for the phase selected (e.g., mentions and concepts) and (D) second pipeline outputs for the phase selected. The mentions/concepts considered for determining the report labels are regarded as *key* mentions/concepts and are differentiated by the *excluded* ones. Here, 2 concepts are identified in the first pipeline, namely, Colon Hyperplastic Polyp and Severe Colon Dysplasia, while in the second one only Severe Colon Dysplasia has been identified. Nevertheless, Colon Hyperplastic Polyp and Sigmoid colon are negligible concepts (i.e., false positives) both associated with the polyp sigmoid mention. In contrast, Severe Colon Dysplasia is correct since it has been identified using a SKET rule verified by the severe dysplasia key mention.

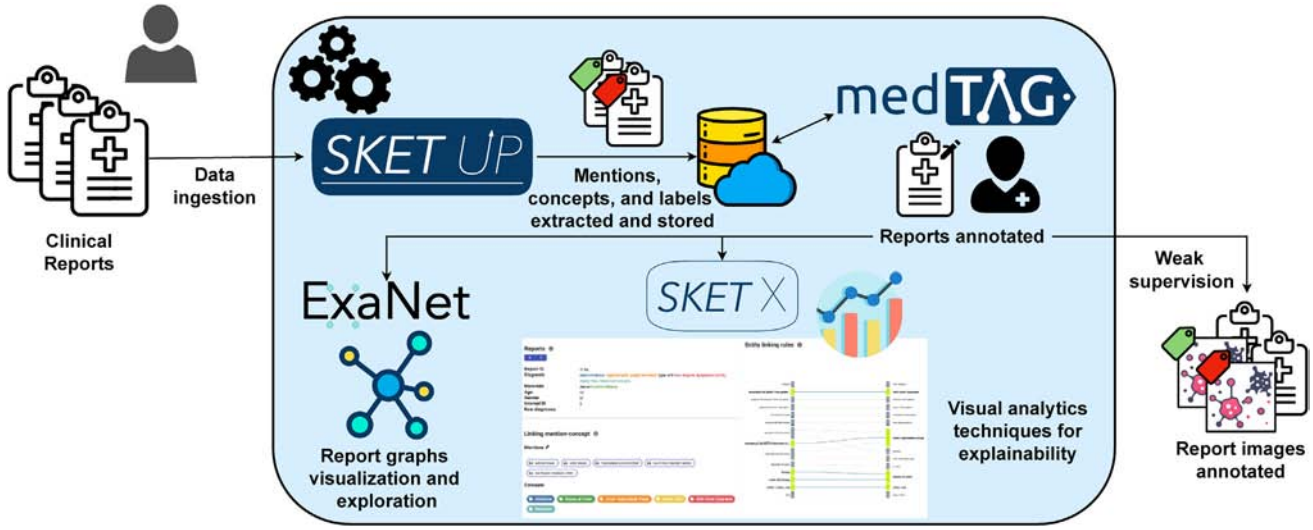


Fig. 8. The SKET ecosystem. From clinical reports, a suite of different applications relying on SKET to process (SKET UP), analyze and annotate (medTAG), explore (ExaNET), and explain (SKET X) the knowledge contained within reports – also providing weak supervision to train cancer-assisted diagnosis tools.

where SKET UP represents the online access point to interact with SKET, SKET X provides explanations for SKET results, medTAG[33] integrates SKET automatic annotations to support semi-automatic tagging, and ExaNet[33] allows to visualize and explore SKET report-level knowledge graphs. Moreover, SKET labels can also be used to supervise cancer-assisted diagnosis tools.[34]

Automatic report annotation

SKET has been integrated as an automatic annotator within MedTAG.[7,33] MedTAG is a collaborative biomedical annotation tool that provides 4 annotation types:

(1.) **Labels:** Allows the user to assign, by clicking on the check-boxes, one or more labels to a document. The labels indicate some reports' properties (e.g. "Cancer" label indicates the presence of a cancer-related disease).
(2.) **Concepts:** Allows the user to specify which concepts are relevant for a document. Users can take advantage of auto-complete functionalities for searching the relevant concepts to assign to each document.
(3.) **Mentions:** Shows the list of the mentions identified by the user in the report text.
(4.) **Linking:** Allows the user to link the mentions identified with the corresponding concepts. Users can link the same mention to multiple concepts.

For each annotation type, SKET provides automatic annotations for reports associated with colon, cervix, and lung use-cases. At present, MedTAG has been used by experts to produce more than 7000 annotations. On the other hand, SKET annotations within MedTAG exceed 100 000 units. Table 5 reports SKET annotation statistics for each annotation type.

Pathological knowledge visualization

The report-level knowledge graphs produced by SKET can be explored with ExaNet[8]. ExaNet is a visual application that allows users (e.g., experts and pathologists) to explore the pathology reports linked data by using an interactive graph visualization tool. ExaNet enables users to explore graph connections by leveraging pan and zoom functionalities. On top of this, ExaNet allows users to visualize an interactive JSON serialization of each pathology report, providing also download capabilities.

Table 5
Number of labels, concepts, mentions, and links automatically annotated by SKET within MedTAG. Statistics are reported for each use-case and globally.

Annotation type	Colon	Cervix	Lung	Total
Labels	9309	16 033	2066	27 408
Concepts	11 932	12 936	2336	27 204
Mentions	10 926	12 070	2336	25 332
Linking	11 932	12 936	2336	27 204
Total	44 099	53 975	9074	107 148

Conceptually, ExaNet stems from ontology visualization tools. The visualization of ontologies is a fundamental task to assess ontologies and enable users to explore, verify, and understand them and their underlying structures.[60–63] Nevertheless, compared to ontology visualization, where the focus is primarily on the Terminological Box (TBox) – i.e., definition of classes and properties – ExaNet focuses instead on the Assertional Box (ABox) – i.e., individuals and instance data. Furthermore, ExaNet replaces the classes Internationalized Resource Identifier (IRI) with the corresponding literals.

WSI classification

The labels produced by SKET are used to reduce supervised-training limitations for colon cancer-assisted diagnosis tools[34] – limitations that prevent the full exploitation of digital pathology applications. In other words, SKET labels serve as weak labels to train a deep image classifier.

The proposed model, based on Multiple Instance Learning Convolutional Neural Networks (CNNs), makes multi-class predictions at patch-level and then aggregates them through an attention pooling layer[64,65] to obtain multi-label WSI predictions. The multi-label setting reflects the very nature of the pathology domain, where images (and reports) can highlight multiple findings for the same sample. Therefore, employing models that produce multi-label predictions allows to better approximate real-world pathology scenarios.

The proposed approach has been trained and tested using data composed of colon WSIs from AOEC and RUMC medical centers. The training set consists of the WSIs associated with the 3769 colon reports reported in Table 1, whereas the test set consists of 227 WSIs from AOEC and 423 from RUMC, for a total of 650 WSIs. Colon cancer was chosen as use-case due to its high social impact and difficulty in diagnosing it. In fact, colon

[7] MedTAG is available at https://github.com/MedTAG/medtag-core/.
[8] ExaNet can be accessed through the "Reports' stats" functionality of MedTAG, under the "Graph" feature associated with each report that has been annotated by SKET.

Table 6
CNN colon cancer performance when trained with SKET weak labels (CNN-SKET) and with manual ones (CNN-GT). Results refer to WSI classification on AOEC and RUMC data. For each considered measure, we report the average obtained through 10-fold cross-validation. **Bold** values represent the highest scores achieved for each measure.

Model	Accuracy	Micro F1	Weighted F1
CNN-SKET	0.6666	0.7741	0.7694
CNN-GT	**0.6795**	**0.7866**	**0.7800**

cancer is the fourth most diagnosed cancer in the world.[66] Besides, the need to identify malignant polyps – which are cell agglomerations protruding from the colon surface – makes it problematic to diagnose.[66] Thus, to prove the effectiveness of SKET as a weak annotator, we compared the performance of the image classifier trained with SKET labels against its performance when trained using manual labels. Table 6 reports the results for subset accuracy, micro-, and weighted-average F1 measures, obtained through 10-fold cross-validation.

The obtained results show the effectiveness of SKET when used as a weak annotator. The performance obtained using weak labels are close to those achieved with manual ones. Precisely, the performance difference between the 2 CNNs does not exceed 1.3%. Furthermore, we performed the Wilcoxon Rank-Sum test and verified that such performance difference is not statistically significant (p-value < 0.05). Thus, SKET allows training cancer diagnostics models for digital pathology without human intervention, paving the way to the use of ML models in the clinical practice.

Conclusions and future work

In this work, we presented the Semantic Knowledge Extractor Tool (SKET), an unsupervised hybrid knowledge extraction system that combines rule-based techniques with pre-trained Machine Learning (ML) models to extract critical pathological concepts from diagnostic reports. The concepts extracted from diagnostic reports can serve different digital pathology applications, such as automatic annotation, knowledge visualization, discovery, or image classification. A throughout evaluation demonstrated SKET effectiveness in annotating colon, cervix, and lung cancer use-cases – making it a viable solution to reduce pathologists' workload. The results and analyses highlighted the importance of expert knowledge in developing unsupervised systems for specialized medicine. Moreover, the effectiveness of SKET as a weak annotator suggests that it can be used as a first, cheap solution to bootstrap supervised models in the absence of manual annotations.

Together with SKET, we also introduced SKET X, a web-based system to support pathologists and domain experts in the understanding of SKET outputs, as well as the role that its different components have on such outputs. Through SKET X, users can comprehend predictions and get valuable insights into the knowledge extraction process. Beyond explainability, SKET has also been used to empower different digital pathology downstream applications. In particular, SKET labels have been used to reduce training limitations for colon cancer-assisted diagnosis tools. The use of SKET for training deep image classifiers without human intervention paves the way to ML models in the clinical practice.[34]

As future work, we plan to extend SKET to other emerging but under-researched use-cases, such as celiac disease – whose prevalence has significantly increased over the past 20 years.[67]

Contributions

S.M. and G.S. designed the experiments, S.M. designed SKET, wrote the code, performed the experiments, analyzed the results for SKET, F.G. and G.S. designed SKET X, F.G. wrote the code, tested, and refined SKET X, N.M. contributed to the analysis of SKET components and is the main contributor of the image classification algorithm using SKET labels, M.A. is the P.I. of the ExaMode project, contributed to design the project and to the ideas behind SKET, S.B. contributed to design the project, to the design of the ExaMode ontology and to the refinement of SKET extracted concepts, G.B. contributed to the definition of the disease classification labels, F.C. contributed to the definition of the disease classification labels, provided ideas for medTAG, the image classification algorithm and helped with the refinement of SKET, G.M.D.N. contributed to the multilingual aspect of SKET and the translation algorithms, F.F. provided the essential knowledge about digital pathology and the considered diseases, O.I. contributed to the development of medTAG, H.M. is one of the main contributors of the ExaMode project and contributed to design the work, T.P. contributed to design the project, to the design of the ExaMode ontology and to the refinement of SKET extracted concepts, S.V provided the knowledge about considered diseases and provided the data used by SKET, G.S. contributed to design the project and supervised all the phases of the work. All authors contributed to the preparation and revision of the manuscript.

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Conflict of Interest

Filippo Fraggetta is an author of this work and a member of the editorial board.

Acknowledgments

Gianmaria Silvello reports financial support was provided by University of Padua. Gianmaria Silvello reports financial support was provided by the European Commission ExaMode project, as part of the EU H2020 program under Grant Agreement no. 825292.

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6:94–98.
2. Buckley JM, Coopey SB, Sharko J, et al. The feasibility of using natural language processing to extract clinical information from breast pahology reports. J Pathol Inform 2012;3:23.
3. Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. Artif Intell Med 2016;66:29–39.
4. Burger G, Abu-Hanna A, de Keizer N, Cornet R. Natural language processing in pathology: a scoping review. J Clin Pathol 2016;69:949–955.
5. Topaz M, Murga L, Gaddis KM, et al. Mining fall-related information in clinical notes: comparison of rule-based and novel word embedding-based machine learning approaches. J Biomed Inform 2019;90.
6. Oliwa T, Maron SB, Chase LM, et al. Obtaining knowledge in pathology reports through a natural language processing approach with classification, named-entity recognition, and relation-extraction heuristics. JCO Clin Cancer Informatics 2019:1–8.
7. Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Informatics 2017;73:14–29.
8. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. J Biomed Informatics 2018;77:34–49.
9. Santus E, Schuster T, Tahmasebi AM, et al. Exploiting rules to enhance machine learning in extracting information from multi-institutional prostate pathology reports. JCO Clin Cancer Informatics 2020:865–874.
10. Kim Y, Lee JH, Choi S, et al. Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. Sci Rep 2020:1–9.
11. Giannaris P, Al-Taie Z, Kovalenko M, et al. Artificial intelligence-driven structurization of diagnostic information in free-text pathology reports. J Pathol Informatics 2020;11:10.
12. Gregg JR, Lang M, Wang LL, et al. Automating the determination of prostate cancer risk strata from electronic medical records. JCO Clin Cancer Informatics 2017:1–8.
13. Glaser AP, Jordan BJ, Cohen J, Desai A, Silberman P, Meeks JJ. Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. JCO Clin Cancer Informatics 2018:1–8.
14. Roberts K, Demner-Fushman D, Voorhees EM, et al. Benchmarking information retrieval for precision oncology: the TREC precision medicine track. AMIA 2018, American

Medical Informatics Association Annual Symposium, San Francisco, CA, November 3-7, 2018, AMIA; 2018.
15. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal 2019;54:280–296.
16. Aeffner F, Wilson K, Martin NT, et al. The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth. Archiv Pathol Lab Med 2017;141:1267–1275.
17. del Toro OJ, Otálora S, Andersson M, et al. Analysis of histopathology images: from traditional machine learning to deep learning. Biomedical Texture Analysis. Academic Press; 2017. p. 281–314.
18. Bejnordi BE, Veta M, Diest PJV, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199–2210.
19. Schaumberg AJ, Juarez W, Choudhury SJ, et al. Large-scale annotation of histopathology images from social media. BioRxiv 2018:1-26. https://doi.org/10.1101/396663.
20. Courtiol P, Tramel EW, Sanselme M, Wainrib G. Classification and disease localization in histopathology using only global labels: a weakly-supervised approach. CoRR 2018:1-13. https://doi.org/10.48550/ARXIV.1802.02212.
21. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis, computational and structural. Biotechnol J 2018;16:34–42.
22. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301–1309.
23. Carbonneau MA, Cheplygina V, Granger E, Gagnon G. Multiple instance learning: a survey of problem characteristics and applications. Pattern Recognit 2018;77:329–353.
24. Dhrangadhariya A, Otálora S, Atzori M, Müller H. Classification of noisy free-text prostate cancer pathology reports using natural language processing. ICPR, Artificial Intelligence for Digital Pathology Workshop (AIDP); 2021.
25. Chiticariu L, Li Y, Reiss FR. Rule-based information extraction is dead! long live rule-based information extraction systems!. Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, ACL; 2013. p. 827–832.
26. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Proc. of the 27th Annual Conference on Neural Information Processing Systems 2013, NIPS, Lake Tahoe, Nevada, United States, December 5-8, 2013; 2013. p. 3111–3119.
27. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguistics 2017;5:135–146.
28. Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, ACL; 2018. p. 2227–2237.
29. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR 2018:1-11. https://doi.org/10.48550/ARXIV.1810.04805.
30. Wang F, Preininger AM. Ai in health: state of the art, challenges, and future directions. Yearb Med Informatics 2019;28:16–26.
31. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Mining Knowl Discov 2019;9.
32. Holzinger A. From machine learning to explainable ai. 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA); 2018. p. 55–66.
33. Giachelle F, Irrera O, Silvello G. MedTAG: a portable and customizable annotation tool for biomedical documents. BMC Med Inform Decis Making 2021;21:352.
34. Marini N, Marchesin S, Otálora S, et al. Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations. npj Digit Med 2022;5.
35. Srigley JR, McGowan T, Maclean A, et al. Standardized synoptic cancer pathology reporting: a population-based approach. J Surg Oncol 2009;99:517–524.
36. Ellis DW, Srigley J. Does standardised structured reporting contribute to quality in diagnostic pathology? the importance of evidence-based datasets. Virchows Arch 2016;468.
37. Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, et al. Marian: fast neural machine translation in C++. Proc. of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations, ACL; 2018. p. 116–121.
38. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Proc. of the 30th Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA; 2017. p. 5998–6008.
39. Marchesin S. Case-based retrieval using document-level semantic networks. Proc. of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, ACM; 2018. p. 1451.
40. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. Proc. of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019, ACL; 2019. p. 319–327.
41. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text processing. Proc LBM 2013:39–44.
42. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Informatics 2001;34:301–310.
43. Ratcliff JW, Metzener DE. Pattern matching: the gestalt approach. Dr Dobbs J 1988;13:46.
44. Shaw JA, Fox EA. Combination of multiple searches. Proc. of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, volume 500-225 of NIST Special Publication, NIST; 1994. p. 105–108.
45. Agosti M, Marchesin S, Silvello G. Learning unsupervised knowledge-enhanced representations to reduce the semantic gap in information retrieval. ACM Trans Inf Syst 2020;38:1-48.
46. Chang WC, Yu HF, Zhong K, Yang Y, Dhillon IS. Taming pretrained transformers for extreme multi-label text classification. KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, ACM; 2020. p. 3163–3171.
47. Ruas P, Andrade VDT, Couto FM. Lasige-biotm at MESINESP2: entity linking with semantic similarity and extreme multi-label classification on spanish biomedical documents. Proc. of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of CEUR Workshop Proceedings; 2021. p. 324–334 CEUR-WS.org.
48. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scient Data 2019;6:1–9.
49. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. CoRR 2019:1–7. https://doi.org/10.48550/ARXIV.1904.03323.
50. Marchesin S, Silvello G. TBGA: a large-scale gene-disease association dataset for biomedical relation extraction. BMC Bioinform 2022;23:111.
51. Dzobo K, Adotey S, Thomford NE, Dzobo W. Integrating artificial and human intelligence: a partnership for responsible innovation in biomedical engineering and medicine. Omics J Integr Biol 2020;24(5):247–263.
52. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3.
53. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. Yearb Med Informatics 2019;28:128–134.
54. Montani S. Exploring new roles for case-based reasoning in heterogeneous AI systems for medical decision support. Appl Intell 2008;28:275–285.
55. Holzinger A. Explainable AI and multi-modal causability in medicine. i-com 2021;19:171–179.
56. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? CoRR 2017:1-28. https://doi.org/10.48550/ARXIV.1712.09923.
57. Thomas JJ, Cook KA. *Illuminating the path: the research and development agenda for visual analytics*. 2005.
58. Angelini M, Santucci G, Schumann H, Schulz H. A review and characterization of progressive visual analytics. Informatics 2018;5:31.
59. Giachelle F, Silvello G. A progressive visual analytics tool for incremental experimental evaluation. Proceedings of the 10th Italian Information Retrieval Workshop, Padova, Italy, September 16-18, 2019; 2019. p. 2–5.
60. Lohmann S, Negru S, Haag F, Ertl T. Visualizing ontologies with VOWL. Semantic Web 2016;7:399–419.
61. Lohmann S, Link V, Marbach E, Negru S. Webvowl: web-based visualization of ontologies. Knowledge Engineering and Knowledge Management - EKAW 2014 Satellite Events, VISUAL, EKM1, and ARCOE-Logic, Linköping, Sweden, November 24-28, 2014. Revised Selected Papers, volume 8982 of LNCS, Springer; 2014. p. 154–158.
62. Lohmann S, Negru S, Haag F, Ertl T. VOWL 2: user-oriented visualization of ontologies. Proc. of the 19th International Conference on Knowledge Engineering and Knowledge Management EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings, volume 8876 of LNCS, Springer; 2014. p. 266–281.
63. Lanzenberger M, Sampson J, Rester M. Visualization in ontology tools. 2009 International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2009, Fukuoka, Japan, March 16-19, 2009, IEEE Computer Society; 2009. p. 705–711.
64. Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. Proc. of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of Proc. of Machine Learning Research, PMLR; 2018. p. 2132–2141.
65. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 2021;5(1):555–570.
66. Benson AB, Venook AP, Al-Hawary MM, et al. NCCN guidelines insights: colon cancer, version 2.2018. J Natl Compreh Cancer Network 2018;16:359–369.
67. King JA, Jeong J, Underwood FE, et al. Incidence of celiac disease is increasing over time: a systematic review and meta-analysis. Off J Am Coll Gastroenterol 2020;115:507–525.
68. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, The Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Inform Decis Mak 2020;20(1):310.
69. EU AI HLEG. A Definition of Artificial Intelligence: Main Capabilities and Scientific Disciplines. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines 2019. Accessed September 5, 2022.
70. EU AI HLEG. Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai 2019. Accessed September 5, 2022.