Exploiting biomedical literature to mine out a large multimodal dataset of rare cancer studies

Anjani Dhrangadhariya[a], Oscar Jimenez-del-Toro[a], Vincent Andrearczyk[a], Manfredo Atzori[a], and Henning Müller[a, b]

[a]University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[b]University of Geneva (UNIGE), Geneva, Switzerland

ABSTRACT

The overall lower survival rate of patients with rare cancers can be explained, among other factors, by the limitations resulting from the scarce available information about them. Large biomedical data repositories, such as PubMed Central Open Access (PMC-OA), have been made freely available to the scientific community and could be exploited to advance the clinical assessment of these diseases. A multimodal approach using visual deep learning and natural language processing methods was developed to mine out 15,028 light microscopy human rare cancer images. The resulting data set is expected to foster the development of novel clinical research in this field and help researchers to build resources for machine learning.

Keywords: PubMed Central, rare cancer, multimodal classification, natural language processing

1. INTRODUCTION

The U.S. National Cancer Institute (NCI) at the National Institute of Health (NIH) designates a cancer as "rare" when it affects less than 15 out of 100,000 people per year.[1] Such a low prevalence is a major challenge in the study of rare cancers, frequently resulting in a shortage of the robust clinical models needed for their detection and treatment using image analysis tools.[2] The overall lower survival rate of patients with these types of cancers can then be explained, among other factors, by difficulties in their clinical research due to the scarcity of available information about them.[3] Large biomedical data repositories *i.e.* PubMed Central Open Access (PMC-OA) are freely available to the scientific community*. These large repositories can be exploited to advance clinical assessment of these diseases, namely through knowledge aggregation from multiple high-quality scientific studies.

The data elements from the biomedical and life science journal articles contained in PMC-OA can be either visual, *i.e.* article figures, or textual *i.e.* information from the full-text articles or figure captions. Although multiple approaches have been proposed to classify the articles in PMC-OA using text, classification algorithms for the images of the articles are rare.[4] Among the approaches that have been proposed, the majority have reached only a generic modality classification level or similar, *e.g.* light microscopy images, computed tomography images and similar categories.[5,6] A more detailed curation is required to fully benefit from such data, one that goes further than the general modality classification task and also considers the information contained in the images from the publications.

We expand the scope of previously proposed methods for the curation of full-text journal articles to mine out a large multimodal (images and text) data set of rare human cancer studies. In particular, the most pressing challenges in identifying rare cancer images and articles are assessed in this work through the comparison of both text and state-of-the-art visual machine learning methods. This study aims to pinpoint the advantages and limitations of the approaches to curate the various data elements from PMC-OA journal articles.

Further author information: (Send correspondence to Anjani Dhrangadhariya)
Anjani Dhrangadhariya: E-mail: anjani.dhrangadhariya@hevs.ch
*https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

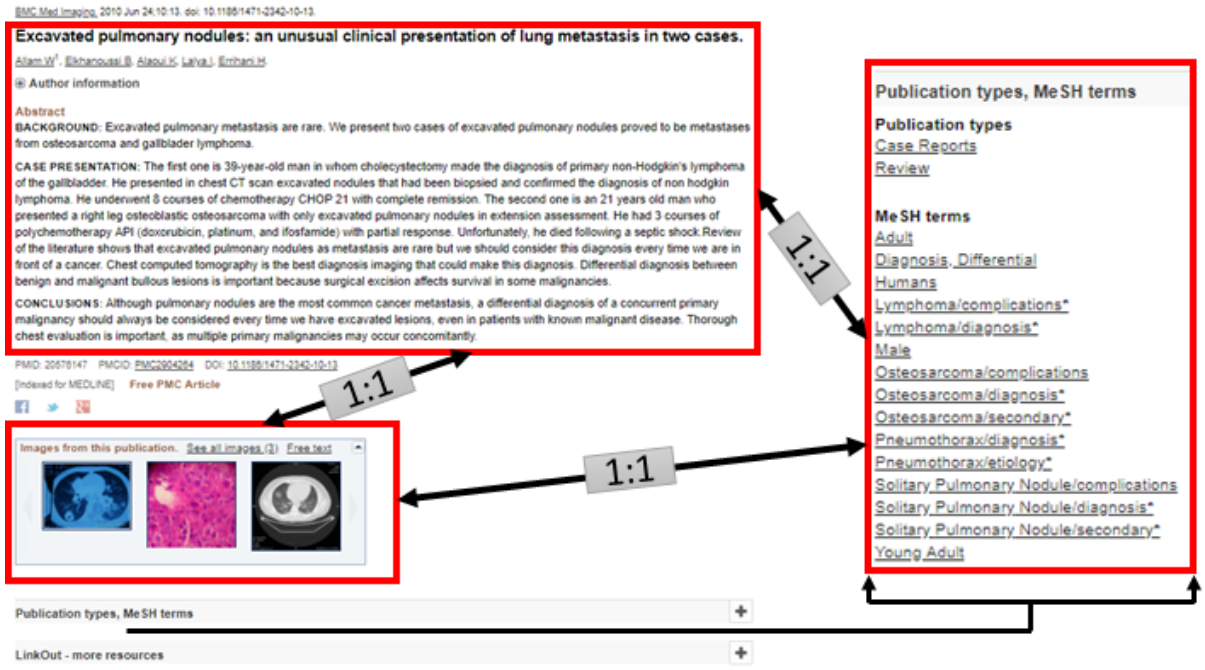

Figure 1: A typical MedLine record explaining the 1:1 association between the journal publication images, the full-text of the article and the manually-annotated MeSH controlled vocabulary.

2. METHODS

The current study is restricted to the curation of Diagnostic Light Microscopy Images (DLMI) from PMC-OA journal articles, as this type of image is fundamental to diagnose cancer cases. Classifying DLMI is a challenging task due to the large variety of staining procedures and slide preparation methods available, as well as the various scale levels of the images.[7] For this work, we initially extract DLMI images from the PMC-OA using a state-of-the-art approach for modality classification, relying only on the images from the publications.[8] An evaluation is then performed for visual and text machine learning approaches for three classification tasks: 1) human vs. non-human 2) neoplastic (tumor-related) vs. non-neoplastic and 3) rare cancer vs. non-rare cancer. The data sets used in the study and the methods developed for each task are explained in this Section.

2.1 Data sets

Medline is one of the largest biomedical citation databases indexing about 30 million biomedical and life science journal publications, online books and their related bibliographic metadata. Medline data can be freely queried and downloaded using the PubMed interface[†]. However, not all the data indexed by Medline, and accessible through PubMed, is available for redistribution and reuse. PubMed Central Open Access (PMC-OA) subset is a free archive for full-text biomedical and life science journal articles, including over 2.09 million publications in 2019. All the publications in the PMC-OA subset are available with CC-BY and CC0 license allowing more liberal redistribution and reuse.[9] An estimated total of 6,736,759 images are present in the full data set of journal publications. The journal publications stored in MedLine are structured as MedLine records that can be accessed using the PubMed interface. A typical MedLine record comprises the publication text, the set of images included in this particular publication as jpg or png files, if available the manually-attached MeSH (Medical Subject Headings) terms, and other metadata.[10,11] Hence, all the data elements within an individual Medline record, share a 1:1 relationship with each other as shown in Fig. 1. MeSH is a hierarchically-organized terminology used for cataloguing the biomedical documents in MedLine[‡] It is organized in a tree structure with

[†]https://www.nlm.nih.gov/bsd/difference.html
[‡]https://www.nlm.nih.gov/mesh/meshhome.html

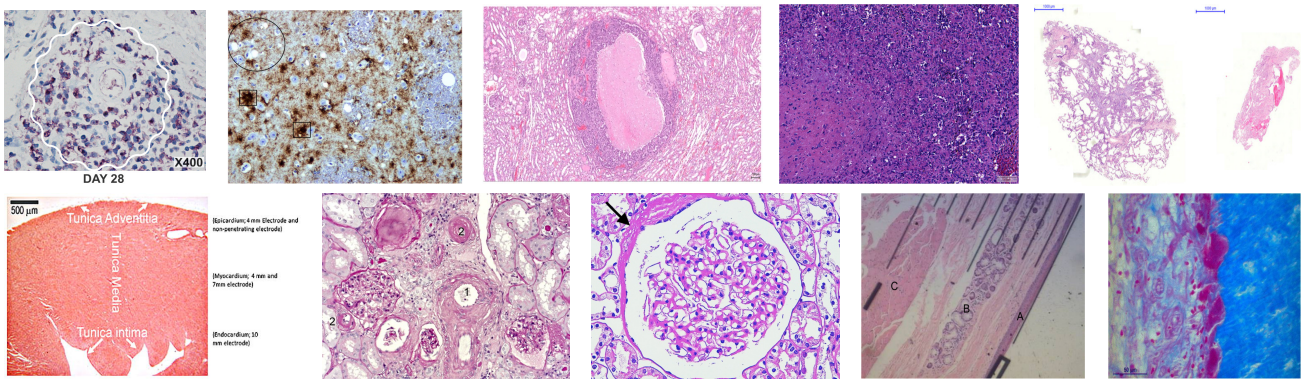

Figure 2: Sample images from the PMC-OA dataset classified as DLMI by the reference method from the ImageCLEF 2018 challenge.

its root node branching into 16 broad thematic categories like "Eukaryota" (organisms), "Diseases", "Chemical and Drugs", *etc.* These categories are further divided into subcategories[§] or MeSH headings. Each MeSH heading in the MeSH tree has its unique MeSH ID and a unique code indicating its location in the tree.[10] It is to be noted that MeSH terms were manually assigned to only a subset of documents in MedLine by the indexers at NLM (National Library of Medicine) and that a large part of MedLine records do not have them. These manually-attached MeSH terms, however, can be considered as the gold standard annotations for the corresponding documents.[11]

2.2 Image classification with convolutional neural networks

In the past few years, deep learning approaches have outperformed classical machine-learning algorithms in many tasks related to the analysis of histopathology images.[12] Unlike methods based on hand-crafted features, deep learning approaches build increasingly abstract representations of the data that are learned in a hierarchical fashion. Moreover, convolutional neural networks (CNN), a commonly used deep learning architecture in image analysis, have shown promising results in the supervised and unsupervised automatic classification of light microscopy images.[13]

As an initial step in this work, a DenseNet-169 convolutional neural network was used to generate a primary modality classification of all the images contained in the PMC-OA journal articles.[14] The approach classified all of the PMC-OA images into 31 image modality types from the ImageCLEF 2018 challenge.[15] The classes include modalities of diagnostic images (*e.g.* DLMI, ultrasound and computed tomography), types of generic biomedical illustrations (*e.g.* tables and flowcharts) and compound figures (images consisting of several sub-figures). The network originally pretrained on ImageNet was fine-tuned with an Adam optimizer and enhanced with data augmentation for this task. More details on the implementation and training setup are given in.[8] Each individual DLMI image classified by the network in this step was then linked to its respective PMC-OA record using the corresponding PMC identifier (PMCID). In Figure 2, a few examples images classified as DLMI by this approach are shown.

To evaluate the performance of a classification strategy, that relies only on the images from the PMC-OA publications, CNN models were then fine-tuned for the selected classification tasks. After the initial modality classification step, the resulting set of DLMIs were further classified with VGG19 networks, trained both with and without data augmentation.[16] Their performance was evaluated for each classification task and compared against text-based approaches.

2.3 Text data preprocessing

Different machine learning methods were built using the text information from the journal articles and the manually-annotated MeSH terms, when available. However, text requires a thorough preprocessing before it can

[§]https://meshb.nlm.nih.gov/treeView

Table 1: List of the corpus-specific stop-words.

introduction	abstract	background	method(s)
materials	objective(s)	aim(s)	outcome(s)
conclusion(s)	result(s)	discussion	methodology
also	may	level(s)	show

be used for any extrinsic natural language processing task. All the MedLine text records were first lowercased before being tokenized into words by Natural Language Toolkit (NLTK)[¶]. The most frequent and noisy tokens were removed using a set of predefined stop words provided by NLTK together with specific stop words from PubMed[‖]. Additional corpus specific stop words were identified during the experiments and the terms were removed accordingly (see Table 1). These stop-words were present in all the texts in the corpus and were assumed to not help the classification process. By any means the list of corpus-specfic stop-words is not complete and could be improved further. Then, a process of text normalization converted British English terms into American English. Lemmatization was performed using the word net lemmatizer.[17] A corpus vocabulary was constructed from all the unique corpus tokens. To scale this vocabulary down, uninformative short tokens with fewer than three characters and also the tokens with vocabulary count lower than five were removed assuming they were not representative of the classes.

2.4 Text representation

Proper representation of text is critical when classifying documents using machine learning methods. Text representation methods convert text documents into a mathematical form or numeric vector understood by machine learning systems. To convert the pre-processed texts from each MedLine record into a numeric vector, three methods for text representation were tested: 1) count-based vectors, 2) word vectors and 3) paragraph vectors.

Count-based representation: One of the earliest count-based representations is the bag-of-words (BoW) representation which involves representing a document by word counts for the words in the document. Term Frequency/Inverse Document Frequency (TF/IDF) is a weighted, count-based method for vectorized text representation. It is a traditional, sparse representation of text and is also a strong baseline.[18] TF-IDF is defined by term frequency (TF) multiplied by inverse document frequency (IDF) as shown in the equation 1. TF measures how frequently a term (t) occurs in a document (d) normalized by the document length. IDF weighs a term based on the number of documents (N) a particular term (t) occurs in divided by the document frequency of that term. This increases the weights for the meaningful words in the corpus and reduces the weights for frequently occurring stop-words like a, an, the, in, if, of, etc.

$$TF - IDF = TF_{t,d} * log(N/DF_t) \tag{1}$$

Word vectors: Word2vec and Global Vectors for Word Representation (GloVe), also called word embeddings, are the two unsupervised algorithms for the extraction of dense, semantic, real-valued vectors from words based on their context.[19,20] While word2vec is a predictive model, GloVe is a count-based model that takes into account the co-occurrence of neighbouring words for the generation of vectors. Both methods capture word semantics, unlike TF-IDF. In the semantic space of a word embedding, vectors for two similar words will be located near each other and have a high cosine similarity. For example, the cosine similarity for the word2vec google news model between the terms "woman" and "patient" is 0.7299, while for the terms "mouse" and "patient" it is 0.3211. For this work, three pre-trained word2vec embeddings, one pre-trained GloVe embedding and a corpus-specific word2vec embedding (see Table 2 for the details) were tested.[19–21] Corpus-specific 300-dimensional word vectors were trained using the word2vec algorithm implemented by Facebooks' fastText.[22] This work considers

[¶]https://www.nltk.org/api/nltk.tokenize.html

[‖]https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/

Table 2: List of word embeddings used in this work. Details are provided on the types of embeddings, the text corpus used to train them and their two characteristic parameters: vector length and window size used to train the embeddings.

Embedding name	Corpus	length and window size (WS)
word2vec	Google news	300-dimensional, WS 5
bio2vec	PubMed	300-dimensional, WS 2
bio2vec	PubMed	300-dimensional, WS 30
GloVe	Wikipedia 2014, English Gigaword	200-dimensional, WS 6
word2vec	Corpus-specific	300-dimensional, WS 5

additive compositionality of word vectors and averages over all the word vectors in a single MedLine study text to get a document-level representation.

Paragraph vectors: Paragraph vectors are generated in an unsupervised manner and learn distributed representation for pieces of text rather than for the individual words. Paragraph vectors learn to associate words with document labels rather than with the other words in context. This work used two kinds of paragraph vectors: 1) a distributed memory model of paragraph vectors (PV-DM) and 2) a distributed bag of words model of paragraph vectors (PV-DBOW).[23]

2.5 Ontology-assisted text classification

All the MeSH-labeled text records were exploited to serially train and evaluate multiple classifiers (Logistic Regression (LR), Support Vector Machines (SVM) with linear kernel and K-nearest neighbor (KNN)) in order to retrieve human rare cancer publications. At each classification step, the classifier performance was evaluated on an independent validation set with the corresponding MeSH terms regarded as the ground truth. GridSearch was used to identify the best text representation, best model and model parameters. Only the setups with the best F1-scores on an unseen validation set were then used to classify the unlabeled text data (without manual MeSH terms).

3. EXPERIMENTAL SETUP

3.1 Mining out "human" images

All the Medline records with images initially classified as DLMI (see section 2.2) were divided into two groups corresponding to the availability of manually-attached MeSH terms. The records with MeSH terms, considered to be the ground truth in this work, were further divided and labeled with three mutually-exclusive labels ("human", "non-human" and "ambiguous"). A MedLine record was labeled "human" if and only if (iff) its corresponding MeSH-term list contained the MeSH code *Humans* [**] (B01.050.150.900.649.313.988.400.112.400.400) and no other organism code from the B01 tree-branch. A record was labeled "ambiguous" iff its corresponding MeSH-term list contained MeSH code *Humans*, but also other organism code from B01 tree-branch, for example, mice[††] (B01.050.150.900.649.313.992.635.505.500) or rat[‡‡] (B01.050.150.900.649.313.992.635.505.700). A record was labeled "non-human" iff no MeSH code *Humans* was present in the MeSH term list. The purpose of this experiment was to precisely retrieve "human" instances. As the "ambiguous" class records mostly represent the animal models of human tissue, this class was merged with the other instances from the class "non-human".

To evaluate the performance of the visual and text approaches in this classification task, the MedLine records with manually-annotated MeSH terms were then divided into independent training, validation and test sets (60-20-20%). The training, validation and test sets used in this and the following tasks were balanced accordingly to have an equal distribution of both classes in each set. Performance measures including F1 score, precision and recall were used.

[**] https://meshb.nlm.nih.gov/record/ui?ui=D006801
[††] https://meshb.nlm.nih.gov/record/ui?ui=D051379
[‡‡] https://meshb.nlm.nih.gov/record/ui?ui=D051381

For the approaches using text, the title and abstract from the selected journal publications were considered. The vectors mentioned in Section 2.4 were extracted from the text corpus in the training set and were used to train and evaluate LR, SVM and KNN models. Exhaustive GridSearch was used to identify the best performing text representation, model and model hyper-parameters.[24]

3.2 Mining out "neoplastic" images

All the images that are MeSH-labeled and predicted "human" in the previous step were carried forward. To obtain the images with a MeSH label "neoplastic", the "human" MeSH-labeled text records were divided into "neoplastic" and "non-neoplastic" based on the presence vs. absence of the MeSH tree code (C04) covering the concept for neoplasms*. This new corpus was again divided into training, validation and test sets (60-20-20%) for this classification task.

For the text approach, the title and abstract from the selected journal publications were considered. The vectors mentioned in Section 2.4 were again extracted from the text corpus in the training set and were used to train and evaluate LR, SVM and KNN models. Exhaustive GridSearch was used to identify the best performing text representation, model and model hyper-parameters.[24]

3.3 Mining out "rare cancer" images

Since there are no manually annotated MeSH terms targeting the concept of "rare cancer", a predefined list of rare cancers provided by the U.S. National Center for Advancing Transnational Sciences (NCATS) at NIH was taken as reference criteria for selecting rare cancer studies†. Only the records classified in the previous tasks as "human" and "neoplastic" were retained as candidate publications for identifying "rare cancer" studies. A keyword-based, string-matching approach was used to filter the text records, extracting only those publications mentioning "rare cancer" terms. The resulting "rare cancer" data set, assembled with text records and their corresponding images, was used as ground truth for assessing the performance of the visual approaches. A convolutional neural network was fine-tuned to differentiate between "rare cancer" images and "non-rare cancer" studies, with and without data augmentation. No text-based classification approach was evaluated for this task, as the text records were used to generate the ground truth.

4. RESULTS

The initial modality classification approach using visual deep learning features classified 241,728 images as "DLMI". Exploiting the 1:1 association (Section 2.1), all the MedLine text records (64,640) and the available MeSH terms corresponding to these "DLMI" images were used. Out of these, a total of 31,733 MedLine text records had manually-attached MeSH terms and 32,907 were unlabeled.

Tables 3 and 4 report the classification performance on the corresponding test sets for the "human" (Section 3.1) and "neoplastic" (Section 3.2) tasks. The best text-based approaches for each classification task reach at least 90% for the F1 score. For the "human" identification step, the TF-IDF tri-gram representation with L2-regularized SVM reached a 90% F1 score. Therefore, the TF-IDF tri-gram with an L2-regularized SVM model was selected as the reference method to classify the unlabeled MedLine records into "human" and "non-human".[23] The wordclouds shown in Figure 3 show similarities between the most frequent terms for Medline text records labelled as "human" (left) and predicted as "human" (right). For instance, the phrases *year old*, *patient*, *case*, *report* and *breast* repeat in both the wordclouds, which qualitatively supports the generated classification model. The wordclouds in Figure 4 show the difference between "human" (left) and "non human" (right) MedLine records identified during the text classification task. In this wordcloud images, prominent phrases like *patients*, *case report*, *year old*, *old male*, and *breast cancer* associate more with "human" group, while the phrases like *stem cell*, *effect*, *observed*, *mouse model* and *treatment* associate more with experiments or studies concerning animal models. As a result of this classification step, a total of 40,314 MedLine text records corresponding to the "human" class were identified. For the DLMI "neoplastic" image identification, the two best approaches reached an F1 close to a perfect classification, with the TF-IDF-bi-grams being slightly better at identifying "neoplastic"

*https://meshb.nlm.nih.gov/record/ui?ui=D009369

†https://rarediseases.info.nih.gov/diseases/diseases-by-category/1

Table 3: Summary and assessment for best visual and text classifier-feature combinations for the "human" classification task, described in Section 3.1.

human vs. non-human classification					
Classifier	Feature type	Feature Settings	Precision	Recall	F1
Visual classification					
VGG19	deep learning	no data augmentation	0.68	0.67	0.67
VGG19	deep learning	with data augmentation	0.69	0.71	0.68
Text classification					
SVM	**count-based**	**TF-IDF tri-grams**	**0.89**	**0.90**	**0.90**
SVM	word vectors	Corpus-specific embeddings	0.88	0.89	0.89
LR	paragraph vectors	PV-DBOW (100, 30, 2)	0.87	0.89	0.88

Figure 3: Word cloud for the 100 most frequent words in MedLine text records labeled "human" vs. the MedLine text records predicted "human" by the experiment in Section 3.1

Figure 4: Word cloud for 100 words most frequent tokens in MedLine text records labeled "human" vs. the MedLine text records predicted "non-human" by the experiment in Section 3.1

Table 4: Summary and assessment for best visual and text classifier-feature combinations for the "neoplastic" classification task, described in Section 3.2 and its comparison to the visual classification

neoplastic vs. non-neoplastic classification					
Classifier	Feature type	Features	Precision	Recall	F1
Visual classification					
VGG19	deep learning	no data augmentation	0.64	0.61	0.63
VGG19	deep learning	with data augmentation	0.68	0.65	0.64
Text classification					
SVM	**count-based**	**TF-IDF bi-grams**	**0.99**	**0.99**	**0.99**
SVM	word vectors	Pretrained bio_NLP vectors	0.98	0.94	0.96
LR	paragraph vectors	PV-DBOW	0.98	0.98	0.98

publications. The wordclouds shown in Figure 5 show similarities between the most frequent terms for Medline text records labelled as "neoplastic" (left) and predicted as "neoplastic" (right). For instance, the phrases *cell carcinoma*, *tumor*, *lymph node* repeat in both the wordclouds, which qualitatively supports the generated classification model. The wordclouds shown in Figure 6 show the difference between the most frequent terms for "neoplastic" (left) vs. "non-neoplastic" (right) MedLine text records identified during the text classification task. Phrases like *cell carcinoma*, *breast cancer*, *cancer cell*, *tumor cell*, and *lymph node* clearly associate with "neoplastic" class.

Figure 5: Word Cloud for the 100 most frequent words in Medline text records labeled "neoplastic" vs. the Medline text records predicted "neoplastic" by the experiment described in Section 3.2

Figure 6: Word cloud for 100 most frequent words in Medline text records labeled "neoplasm" vs. the Medline text records predicted "non-neoplastic" by the experiment in Section 3.2

Table 5: Summary and assessment for the visual classification of "rare cancer" vs. "non-rare cancer" images.

rare cancer vs. non-rare cancer classification					
Classifier	Feature type	Features	Precision	Recall	F1
Visual classification					
VGG19	deep learning	no data augmentation	0.61	0.77	0.68
VGG19	**deep learning**	**with data augmentation**	**0.62**	**0.77**	**0.69**

When compared to the VGG19 deep learning model, relying on visual features from the images, the text approaches obtained a much higher precision and better recall as well (Table 4). As a result of the "neoplastic" vs. "non-neoplastic" classification, a total of 12,738 Medline text records corresponding to the "neoplastic" class were identified. The setup with best F1-score from the two previous classification tasks was used to classify the remaining Medline text records with no manually-attached MeSH terms. By applying a keyword-based filtering to the classified data set, 2,669 Medline text reccords and 15,028 light microscopy human rare cancer open-access images were identified together with their corresponding journal articles containing highly-correlated text information. The visual classification for this final task, using deep learning features with data augmentation, resulted in an F1 score of 0.69.

5. DISCUSSION

To the best of our knowledge, this is the first study targeting the automatic extraction of rare cancer images from journal publications from the biomedical literature. The proposed approach relied on both the visual and text information from the freely available publications in the PMC-OA repository. The final pipeline contained the following steps: 1) mining out "DLMI" images, 2) mining out "human" images, 3) mining out "neoplastic" (tumor-related) images, and finally 4) mining out human "rare cancer" light microscopy images with their corresponding journal articles. The output data set contains a large and heterogeneous collection of rare cancer images and publications that could be useful to train and develop novel approaches for these diseases. When comparing the textual vs. visual classification performance in both the "human" vs. "non-human" task, and the "neoplastic" vs. "non-neoplastic" task, text features performed considerably better compared to the visual features. The TF-IDF bi/tri-gram approach with an SVM classifier was the best approach for both tasks, relying on count-based features. Nevertheless, visual features could correctly classify some test images with a recall of up to 0.71 in the "human" identification task. It is important to note that the class "ambiguous" was merged with the class "non-human", thus influencing the classification results for both visual and text features. For the "neoplastic" vs. "non-neoplastic" classification, the visual features gave a worse performance with a maximum F1 score of 0.64. In the final classification task, the developed textual approach was selected as ground truth, since there are no MeSH terms targeting "rare cancers". On the other hand, the visual approaches in the "rare cancer" vs. "non-rare cancer" task, had better results than in the previous 2 classification tasks, but there is still room for improvement. Some of the images and diseases that were identified in this pipeline are shown in Figure 7.

In this work, we compared the advantages and limitations of the visual classification of the images vs. the text classification of some text elements from the corresponding journal publications (title, abstract). The results show that with the current number of images available, the simpler and more interpretable text mining approaches (TF-IDF) outperform state-of-the-art visual strategies. However, it is important to consider that the classification performed on the individual images from a publication is not on the same level as the classification of the full-text records. With the combination of an initial DLMI classification based only on the visual features, and the subsequent text mining non light microscopy images were excluded from the final data set. The PubMed Central Open Access repository includes more than 2.09 million publications and is continuously being updated with novel full-text biomedical scientific publications, including the latest rare cancer studies. Since the PMC-OA dataset will continue to grow in the following years, this is an initial framework to automatically generate high-quality multimodal data sets that could potentially improve the understanding of these type of diseases. Our future research direction is towards experimenting with the multimodal fusion of visual and textual representations from individual images and journal articles.[25]

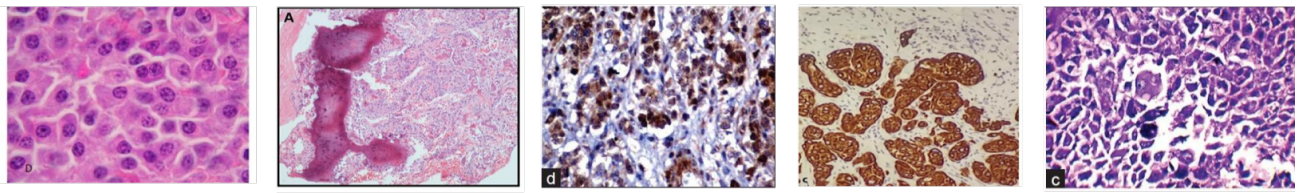

Figure 7: Sample light microscopy human rare cancer images mined out with the proposed approach. Some of the samples come from PMC-OA publications with no MeSH terms available.

6. CONCLUSIONS

A framework that relies on visual deep learning and natural language processing methods to mine out 32,486 light microscopy human rare cancer images is presented. The generated multimodal dataset can fill the void of information regarding the study of these diseases and be further used by researchers as an automatically annotated database for the development of new clinical models. A more comprehensive understanding of the changes present in these cancers, can potentially help to improve the outcome of these patients.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the project ExaMode, grant agreement No 825292.

REFERENCES

[1] Komatsubara, K. M. and Carvajal, R. D., "The promise and challenges of rare cancer research," *The Lancet Oncology* **17**(2), 136–138 (2016).

[2] Puca, L., Bareja, R., Prandi, D., Shaw, R., Benelli, M., Karthaus, W. R., Hess, J., Sigouros, M., Donoghue, A., Kossai, M., et al., "Patient derived organoids to model rare prostate cancer phenotypes," *Nature communications* **9**(1), 2404 (2018).

[3] Gatta, G., Capocaccia, R., Botta, L., Mallone, S., De Angelis, R., Ardanaz, E., Comber, H., Dimitrova, N., Leinonen, M. K., Siesling, S., et al., "Burden and centralised treatment in europe of rare tumours: results of rarecareneta population-based study," *The Lancet Oncology* **18**(8), 1022–1039 (2017).

[4] Müller, H., Andrearczyk, V., Jimenez-del-Toro, O., Dhrangadhariya, A., Schaer, R., and Atzori, M., "Studying public medical images from the open access literature and social networks for model training and knowledge extraction," in [*International Conference on Multimedia Modeling*], 553–564, Springer (2020).

[5] Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., and Müller, H., "Evaluating performance of biomedical image retrieval systemsan overview of the medical image retrieval task at imageclef 2004–2013," *Computerized Medical Imaging and Graphics* **39**, 55–61 (2015).

[6] Kumar, A., Kim, J., Lyndon, D., Fulham, M., and Feng, D., "An ensemble of fine-tuned convolutional neural networks for medical image classification," *IEEE journal of biomedical and health informatics* **21**(1), 31–40 (2016).

[7] Otálora, S., Atzori, M., Andrearczyk, V., and Müller, H., "Image magnification regression using densenet for exploiting histopathology open access content," in [*Computational pathology and ophthalmic medical image analysis*], 148–155, Springer (2018).

[8] Andrearczyk, V. and Müller, H., "Deep multimodal classification of image types in biomedical journal figures," in [*International Conference of the Cross-Language Evaluation Forum for European Languages*], 3–14, Springer (2018).

[9] Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., Connor, R., Fiorini, N., Funk, K., Hefferon, T., et al., "Database resources of the national center for biotechnology information," *Nucleic acids research* **47**(Database issue), D23 (2019).

[10] Lipscomb, C. E., "Medical Subject Headings (MeSH)," *Bull Med Libr Assoc* **88**, 265–266 (Jul 2000).

[11] Mao, Y. and Lu, Z., "MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank," *J Biomed Semantics* **8**, 15 (Apr 2017).

[12] Jimenez-del Toro, O., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rousson, M., Müller, H., and Atzori, M., "Analysis of histopathology images: From traditional machine learning to deep learning," in [*Biomedical Texture Analysis*], 281–314, Elsevier (2017).

[13] Jimenez-del-Toro, O., Atzori, M., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rönnquist, P., and Müller, H., "Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score," in [*Medical Imaging 2017: Digital Pathology*], **10140**, 101400O, International Society for Optics and Photonics (2017).

[14] Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L., "Densely connected convolutional networks," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], (2017).

[15] Ionescu, B., Müller, H., Villegas, M., de Herrera, A. G. S., Eickhoff, C., Andrearczyk, V., Dicente Cid, Y., Liauchuk, V., Kovalev, V., Hasan, S. A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.-T., Piras, L., Riegler, M., Zhou, L., Lux, M., and Gurrin, C., "Overview of ImageCLEF 2018: Challenges, datasets and evaluation," in [*Experimental IR Meets Multilinguality, Multimodality, and Interaction*], *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, LNCS Lecture Notes in Computer Science, Springer, Avignon, France (September 10-14 2018).

[16] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

[17] Liu, H., Christiansen, T., Baumgartner, W. A., and Verspoor, K., "BioLemmatizer: a lemmatization tool for morphological processing of biomedical text," *J Biomed Semantics* **3**, 3 (Apr 2012).

[18] Ko, Y., "A study of term weighting schemes using class information for text classification," in [*SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*], 1029–1030 (Aug 2012).

[19] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J., "Distributed representations of words and phrases and their compositionality.," *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems* **2**, 3111–3119 (Dec 2013).

[20] Pennington, J., Socher, R., and Manning, C. D., "Glove: Global vectors for word representation," in [*Empirical Methods in Natural Language Processing (EMNLP)*], 1532–1543 (2014).

[21] Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., and Ananiadou, S., "Distributional semantics resources for biomedical text processing," in [*Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LBM '13)*], 39–44 (2013).

[22] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T., "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics* **5**, 135–146 (2017).

[23] Le, Q. and Mikolov, T., "Distributed representations of sentences and documents," in [*International conference on machine learning*], 1188–1196 (2014).

[24] Chicco, D., "Ten quick tips for machine learning in computational biology," *BioData mining* **10**(1), 35 (2017).

[25] Arevalo, J., Solorio, T., Montes-y Gómez, M., and González, F. A., "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992* (2017).