Multi-Label Modality Classification for Figures in Biomedical Literature

Athanasios Lagopoulos, Anestis Fachantidis, Grigorios Tsoumakas
School of Informatics,
Aristotle University of Thessaloniki
Thessaloniki, Greece
Email: {lathanag,afa,greg}@csd.auth.gr

***Abstract*—The figures found in biomedical literature are a vital part of biomedical research, education and clinical decision. The multitude of their modalities and the lack of corresponding meta-data, constitute search and information retrieval a difficult task. We present multi-label modality classification approaches for biomedical figures. In particular, we investigate using both simple and compound figures for training a multi-label model to be used for annotating either all figures, or only those predicted as compound by an initial compound figure detection model. Using data from the medical task of ImageCLEF 2016, we train our approaches with visual features and compare them with the standard approach involving compound figure separation into sub-figures. Furthermore, we present a web application for medical figure retrieval, which is based on one of our classification approaches and allows users to search for figures of PubMed Central.**

***Keywords*-biomedical figures; modality classification; multi-label learning; information retrieval; PubMed Central**

I. Introduction

Nowadays, a large amount of biomedical figures is publicly available within open access scientific articles. PubMed Central[1] (PMC), the open access subset of PubMed, contains more than 4 million articles and is growing at a rapid pace; 200,000 articles were added in 2014 alone [1]. The figures of these articles can be retrieved through Web interfaces and APIs along with the full-text. However, the lack of associated meta-data, besides the captions, hinders the fulfillment of richer information needs of biomedical researchers, practitioners and educators. The modality of a figure (e.g. angiography, microscopy), in particular, is a very helpful kind of meta-data for medical retrieval [2], [3], [4].

About 40% of the figures in PMC are compound, comprising two or more sub-figures in a multi-panel format [5]. Figure 1 is an example of a compound figure comprising three sub-figures: one *graph* and two obtained through *magnetic resonance* imaging. The *standard approach* to biomedical figure modality classification first uses a binary model to recognize whether the figure is compound or not. If the figure is simple, then a multi-class model is used to predict its modality. If it is compound, then a figure separation algorithm is first invoked to split it into its constituent sub-figures [6], [7], [8]. Then a multi-class model is used to predict the modality of each sub-figure.

[1]http://www.ncbi.nlm.nih.gov/pmc/

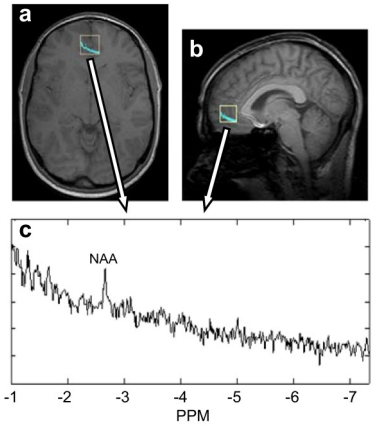

Figure 1. A compound figure comprising 3 sub-figures: one graph (lower part) and two obtained via magnetic resonance imaging (upper part). Adapted from "A review of imaging techniques for systems biology.", Kherlopian, Armen R., et al. BMC systems biology 2.1 (2008): 74.

Figure separation is not perfect. In the last two ImageCLEF benchmarks (2015, 2016), the best figure separation accuracy, which is based on the overlap between predicted and ground truth sub-figures [9], reached approximately 85%. Errors in figure separation propagate to the multi-class model that predicts the modality of the detected sub-figures, harming the overall accuracy in modality classification of compound figures. In addition, classifying sub-figures isolated from their context (the original compound figure they belong to) can lead to information loss, as certain types of modalities might be correlated in compound figures. This is why recently, *multi-label* classification approaches have been investigated for classifying the modalities of compound figures [5], [10], [1], [11].

This paper focuses on using multi-label learning for recognizing the image modalities that characterize a (potentially compound) biomedical figure. It empirically compares the standard approach, based on figure separation, with a number of multi-label learning approaches encompassing novel design elements, such as: (i) using both simple and compound figures for training a multi-label model, and (ii) doing without an initial compound figure detection model. Finally, this paper describes a publicly available information

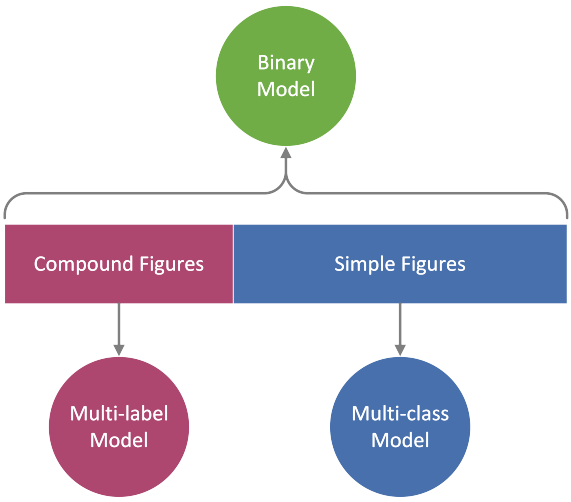

Figure 2. Training with the standard multi-label approach.

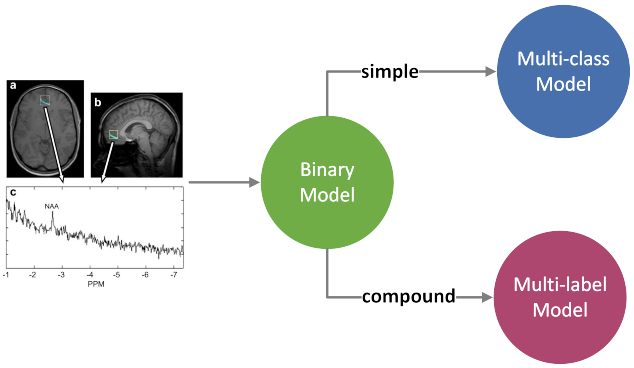

Figure 3. Prediction with the standard multi-label approach.

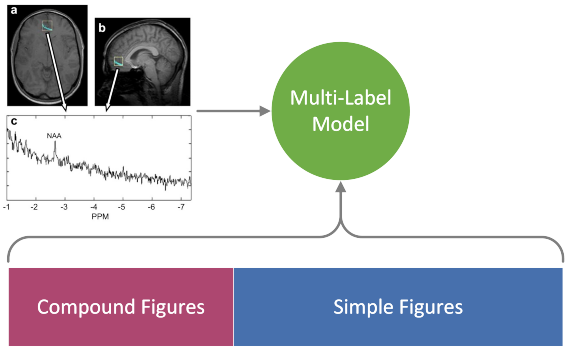

Figure 4. Training and prediction with the simple multi-label approach.

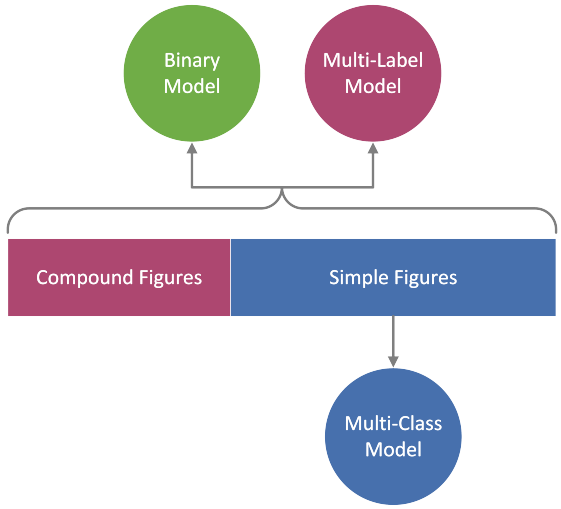

Figure 5. Training with the extended multi-label approach.

retrieval system that runs on top of PMC and incorporates the best model discovered via our empirical work.

The rest of the paper is organized as follows. Section II presents our multi-label learning approaches and outlines their differences from past work. Section III describes the available data and discusses the evaluation of our approaches. Our medical figure retrieval system is described in Section IV and final conclusions along with future work are drawn in Section V.

II. Multi-Label Modality Classification

At first glance, employing multi-label learning for classifying biomedical figures by modality appears simple. The main idea is to change the compound figure classification part of the standard approach from using a figure separation module followed by a multi-class model to using a multi-label model trained on compound figures. Training and prediction with this *standard multi-label* approach is depicted in Figures 2 and 3 respectively.

However, this simple approach forgoes the use of simple figures as training examples for the multi-label model. After all, simple figures can be considered as multi-label training examples annotated with only one label. On one hand, simple figures are fundamentally different from compound ones and putting them all in the same model might require a more complex learning model. On the other hand, exploiting the simple figures for training the multi-label model can increase its discrimination capability for modalities that are under-represented in the set of compound figures.

Therefore, a different and simpler approach we consider here is to learn a single multi-label model from both compound and non-compound figures, essentially treating the latter as multi-label examples with just a single label. Training and prediction with this *simple multi-label* approach is depicted in Figure 4.

Another approach we consider here concerns replacing the multi-label model of the standard multi-label approach with that of the simple multi-label approach. In other words, we use both compound and non-compound figures for training the multi-label model in the standard multi-label approach. Training with this *extended multi-label* approach is depicted in Figure 5. Prediction remains the same, as depicted in

Figure 3: only figures predicted by the binary model as compound are passed on to the multi-label model.

Past work on multi-label classification of biomedical figures by modality sprung out of the medical task of ImageCLEF 2015 and 2016, where multi-label classification of compound figures was introduced as a sub-task [5], [1]. In particular, two groups participated in this sub-task in 2015 and another two groups in 2016 (one of them was our group). All these works [5], [10], [1], [11] focused solely on multi-label classification of compound figures and did not look at the alternative architectures we discuss here, which take into consideration non-compound figures too, in order to address the full real-world problem of classifying figures (be they compound or non-compound) from open-access biomedical literature. None of these works attempted a comparison with the standard process of figure separation followed by the invocation of a multi-class model per sub-figure.

III. Empirical Study

This section initially describes the data we used for our empirical study and gives details about the evaluation process we followed and about the base learning algorithms we used underneath our approaches. It then presents and discusses the evaluation results of the different approaches for classifying biomedical figures by modality.

A. Data

We experimented with the development set distributed for the medical task of ImageCLEF 2016[2]. The data contain 20,985 figures in JPEG format, of which 12,338 (59%) are compound and 8,647 (41%) non-compound. 1,568 figures (13%) of the set of compound figures are further annotated with one or more classes out of a hierarchy of 30 modality classes (see Figure 6), which refer to types of diagnostic images (radiology, visible light photography, printed signals/waves, microscopy, 3D reconstructions) and biomedical illustrations. This hierarchy is a minor extension of the hierarchy used in [12], where a class representing compound figures was also present. Finally, the organizers of the medical task of ImageCLEF 2016 further provide the 6,776 sub-figures of these 1,568 compound figures, along with their annotation with one of the 30 classes of the hierarchy of Figure 6.

As we mentioned in Section I, about 40% (60%) of the available figures in PMC are compound (simple). However, the medical task of ImageCLEF 2016 did not provide annotations for the 8,647 non-compound figures it delivered. Therefore, in order to simulate a training set following the distribution of PMC, we adopt the following process. We work with the 1,568 compound figures only, keeping 40% of them as they are and replacing the rest of them with their respective sub-figures, which assume the role of non-compound figures.

[2]http://www.imageclef.org/2016/medical

We extract visual features from the JPEG file of each compound and simulated non-compound figure using the Caffe framework [13]. For each figure, 4,096 features were extracted via the *fc7* (inner product or fully connected) level. Every feature is a non-negative number with 9 decimals. We used the BVLC CaffeNet Model[3], which is a replication of the model described in [14] and has been trained with 1.2 million high-resolution images.

B. Learning Algorithms and Evaluation Process

We employ linear support vector machines (SVMs) from the scikit-learn library[4] to learn all models [15]. We use default parameter settings (cost parameter equal to 1, squared hinge loss function, $L2$ penalization). The one-vs-rest transformation was used to decompose the multi-class learning problem into multiple binary classification tasks, and similarly the binary relevance transformation was used to decompose the multi-label learning problem [16].

All approaches were evaluated using 10-fold cross validation. We first split the 1,568 figures into 10 equally sized disjoint subsets and then we apply the process discussed in the 2nd paragraph of Section III-A separately at each fold. This ensures that sub-figures of the same figure stay within the same fold in order to avoid information leakage from a training to a test set. We use the approach described in [17] to avoid biased cross-validated results.

We use micro- and macro-averaging to compute binary evaluation metrics, such as recall, precision and f-measure, across all classes (labels) in the multi-class (multi-label) tasks. Micro-averaging calculates metrics globally by counting all true positives, false negatives and false positives, while macro-averaging calculates metrics per class/label and then takes the mean across all classes/labels. In the multi-label task, we further use samples-averaging, which calculates metrics per instance and then takes the mean across all instances.

C. Results

Table I shows the micro-, macro- and samples-averaged F-measure for the standard, simple and extended multi-label approaches and for the standard approach assuming perfect figure separation into sub-figures.

We first notice that the simple multi-label approach is the worst of all, highlighting the importance of having a compound figure detection model. The standard and extended multi-label approaches have similar micro- and samples-averaged F-measure, but the extended one leads to slightly higher macro-averaged F-measure. These results appear to be in alignment with our hypothesis that using additional training examples can boost the discrimination capability in rare modality classes, as macro-averaging treats all labels

[3]https://git.io/v484G

[4]http://scikit-learn.org/

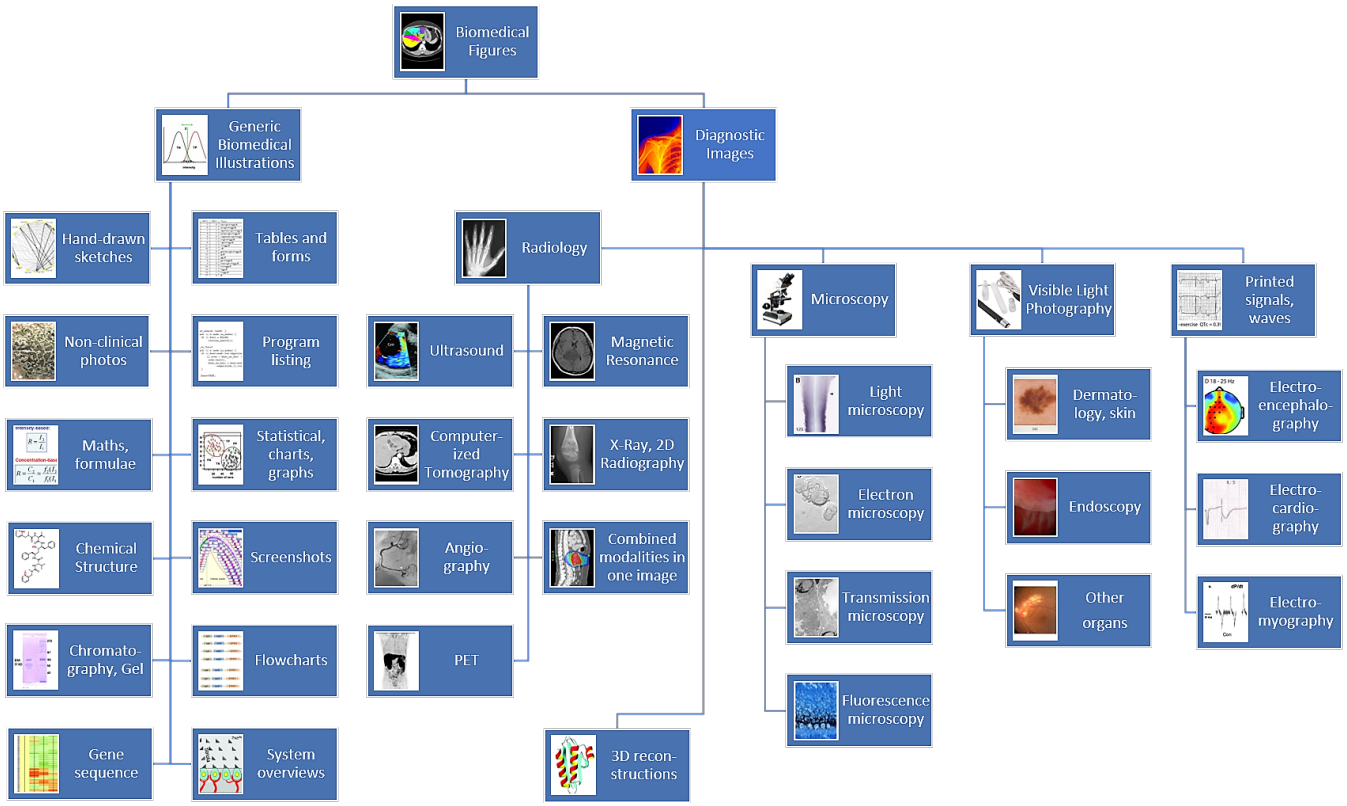

Figure 6. Hierarchy of 30 modality classes concerning various types of diagnostic images and generic biomedical illustrations.

equivalently, while the contribution of each label in micro- and samples-averaging is proportional to its frequency.

To further look into this issue, we studied the per class F-measure in the *standard multi-label* and *extended multi-label* approaches. We did not find consistent improvements across classes, as for 8 (12) classes the F-measure was better with the standard (extended) multi-label approach, and for 10 classes the F-measure did not change between approaches. In 7 out of the latter 10 classes the F-measure was actually zero and this was due to the very small number of training examples (less than 5 examples).

We looked at potential correlations that could explain the per class difference in F-measure between the two approaches, with the first hypothesis concerning the frequency of the classes. However, we found only a negligible correlation ($r \approx -0.1$). The strongest correlation we found, still only a small one ($r \approx -0.35$), concerned the F-measure of the standard approach. This appears to be intuitively meaningful, as the more difficult the learning task (low standard multi-label F-measure), the higher the potential for improvement by obtaining additional training examples through the extended multi-label approach. In the process, we also computed another quantity: the mean cardinality of the multi-label examples of each class, where cardinality is the number of different labels of a multi-label example

Table I
RESULTS COMPARING THE FOUR APPROACHES

Approach \ F-measure	Macro	Micro	Samples
Standard	0.3569	0.7786	0.7912
Standard multi-label	0.3270	0.7667	0.7726
Simple multi-label	0.3139	0.7581	0.7215
Extended multi-label	0.3309	0.7666	0.7728

[16]. Intuitively, this should be correlated with the difficulty of recognizing a class, as the higher its mean cardinality, the higher the number of sub-figures of different modality appear in the same figure. Indeed mean cardinality has a moderate correlation ($r \approx -0.6$) with the F-measure of a class. Table II shows frequencies, f-measures and mean cardinality for the 5 classes with highest improvement (upper part) and deterioration (lower part) of the F-measure when switching from the standard to the extended multi-label approach.

We also notice that the *standard approach* achieves the best results in all measures. However, the differences with the extended multi-label approach are quite small. This is very encouraging for the extended multi-label approach, given that we assumed a perfect separation of compound

Table II
TOP-5 IMPROVED AND TOP-5 DETERIORATED CLASSES WHEN SWITCHING FROM THE STANDARD (STD) TO THE EXTENDED (EXT) MULTI-LABEL APPROACH

Class	Card	F-measure		Frequency	
		Std	Ext	Std	Ext
Screenshot	1.8	0.1571	0.2024	7.6	23.4
Combination	2.29	0.19	0.22	4.9	27.9
Other Organs	2.1	0.1244	0.1431	11.5	44.6
System Overview	2.31	0.1761	0.1994	24.7	70.2
Gene sequence	2	0.3495	0.3816	36.8	124.3
Ultrasound	1.75	0.3833	0.3333	7.1	18.5
Endoscopy	1.91	0.32	0.2867	6.1	12.6
Hand-drawn	2.11	0.3258	0.3147	27.1	95
Transmission	2.01	0.5516	0.5401	38.2	210.9
Tomography	2.06	0.3817	0.3765	14.6	36.5

Table III
RESULTS FOR THE COMPOUND FIGURE DETECTION MODEL

Metric \ Class	Compound	Simple
Recall	0.8016	0.9749
Precision	0.8306	0.9697
F-measure	0.8158	0.9723
Balanced Accuracy	0.8883	
G-mean	0.9407	

Table IV
RESULTS FOR THE MULTI-CLASS MODEL

Metric	Macro	Micro
Recall	0.3958	0.7954
Precision	0.4212	
F-measure	0.4081	

figures into sub-figures. We must also consider that the multi-label models were built with the basic binary relevance approach that treats labels independently. Much more elaborate approaches exist in the literature that take label relationships into account and lead to improved results compared to binary relevance [16].

Table III shows per class recall, precision and F-measure of the binary compound figure detection model, common in the *standard*, *standard multi-label* and *extended multi-label* approaches. The last two rows of the table show the balanced accuracy and the G-mean of this model. We notice that this model is quite accurate and this offers additional evidence in favor of using such an initial model for modality classification of biomedical figures.

For completeness, Table IV shows average recall, precision and F-measure of the multi-class model common in the *standard*, *standard multi-label* and *extended multi-label* approaches, estimated based on the simple figures of our data set (sub-figures of the 60% of the compound figures).

IV. THE MEDIEVAL SYSTEM

This section describes the Web application we have developed for MEDical figure retrIEVAL, dubbed MEDIEVAL[5]. Users can search for PMC figures by entering a text query to be matched against the caption of each figure. MEDIEVAL allows filtering the results by modality, by letting the users select the modalities they are interested in. Figures are sorted according to the similarity of their caption with the text query. Users can see the image and caption of each retrieved figure and navigate to the PMC article containing it. The front-end of MEDIEVAL has been developed with the AngularJS[6] JavaScript framework.

MEDIEVAL retrieves articles from PMC using the PMC-OAI[7] service and extracts the figures. For each figure, it first extracts visual features and then classifies the modality using the extended multi-label approach, with constituent models trained on the full training set of 1,568 compound figures except for the binary classification model. The latter model was trained on the 20,985 figures data set, primarily because it is much larger and secondarily because PMC consists of actual non-compound figures and not sub-figures. The modality predictions (ground truth for the training set) along with the figure's caption, unique PMC ID, URL and useful information about the corresponding articles (i.e., the article's unique PMC ID, title and URL) are stored in a Solr search platform[8] that powers the back-end of our system. MEDIEVAL visits PMC weekly to retrieve new articles.

V. CONCLUSION AND FUTURE WORK

This work discussed the use of multi-label learning models in the modality classification task of figures found in biomedical literature. We investigated using both simple and compound figures for training a multi-label model to be used for annotating either all figures, or only those predicted as compound by an initial compound figure detection model. The proposed approaches allow for a richer modeling of the modality classification task, which not only addresses information loss when treating compound figures as multiple independent figures, but also addresses model redundancy due to building separate models to classify the same underlying modalities. The empirical study of these approaches and their comparison with the compound figure separation approach was based on data from the ImageCLEF 2016 medical task and on well-established evaluation measures and process. The *extended multi-label* approach showed particularly promising results, only slightly worse than using a perfect figure separation approach. Finally, we implemented a Web application, which incorporates the proposed approaches and allows users to search for PMC figures of their preferred modality by caption.

[5]http://atypon.csd.auth.gr/medieval/
[6]https://angularjs.org/
[7]https://www.ncbi.nlm.nih.gov/pmc/tools/oai/
[8]http://lucene.apache.org/solr/

In the future, we plan to investigate how textual features fare in the task of classifying biomedical figures by modality, both by themselves and in tandem with visual features. Towards this, we have already experimented with multigram representations of a figure's caption and of the text referring to the figure within the article, which we found to slightly improve the results compared to using captions alone [1]. We achieved 88.13% accuracy in the compound figure detection sub-task of the ImageCLEF 2016 medical task, which was the best result for a textual only approach. We also achieved 0.32 F-measure in the multi-label classification sub-task, which was equal to the results of the sole other group that participated in this sub-task using a visual approach based on deep learning and convolutional neural networks [11].

Our future plans also include an extension of MEDIEVAL towards taking into account user feedback, to allow for the crowdsourcing of ground truth data, which can then be used to improve the underlying models. Users will be able to give feedback for particular figures returned to them during their search sessions, but we will also employ active learning techniques to explicitly request the feedback that will mostly benefit the system. We will further add a gamification component with a leader-board of weekly/monthly top contributors, which is expected to lead to increased user engagement.

ACKNOWLEDGMENT

This work was partially funded by Atypon Systems Inc.

REFERENCES

[1] A. García Seco de Herrera, R. Schaer, S. Bromuri, and H. Müller, "Overview of the ImageCLEF 2016 Medical Task," in *Working Notes of CLEF 2016 (Cross Language Evaluation Forum)*, Évora, Portugal, 2016.

[2] J. Kalpathy-Cramer and W. Hersh, "Automatic image modality based classification and annotation to improve medical image retrieval." *Studies in health technology and informatics*, vol. 129, no. Pt 2, pp. 1334–8, 2007.

[3] P. Tirilly, K. Lu, X. Mu, T. Zhao, and Y. Cao, "On modality classification and its use in text-based image retrieval in medical databases," in *Proceedings - International Workshop on Content-Based Multimedia Indexing*, 2011, pp. 109–114.

[4] D. Markonis, M. Holzer, S. Dungs, A. Vargas, G. Langs, S. Kriewel, and H. Müller, "A survey on visual information search behavior and requirements of radiologists," *Methods of Information in Medicine*, vol. 51, no. 6, pp. 539–548, 2012.

[5] A. García Seco de Herrera, H. Müller, and S. Bromuri, "Overview of the ImageCLEF 2015 medical classification task," in *Working Notes of CLEF 2015 (Cross Language Evaluation Forum)*, Toulouse, France, 2015.

[6] E. Apostolova, D. You, Z. Xue, S. Antani, D. Demner-Fushman, and G. R. Thoma, "Image retrieval from scientific publications: Text and image content processing to separate multipanel figures," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 5, pp. 893–908, 2013.

[7] A. Chhatkuli, A. Foncubierta-Rodríguez, D. Markonis, F. Meriaudeau, and H. Müller, "Separating compound figures in journal articles to allow for subfigure classification," in *Proc. SPIE 8674, Medical Imaging 2013: Advanced PACS-based Imaging Informatics and Therapeutic Applications*, M. Y. Law and W. W. Boonn, Eds., mar 2013, p. 86740J.

[8] K. Santosh, S. Antani, and G. Thoma, "Stitched Multipanel Biomedical Figure Separation," in *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*. IEEE, jun 2015, pp. 54–59.

[9] A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner-Fushman, S. K. Antani, and H. Müller, "Overview of the imageclef 2013 medical tasks." in *CLEF (Working Notes)*, 2013.

[10] A. J. Rodríguez-Sánchez, S. Fontanella, J. Piater, and S. Szedmak, "IIS at ImageCLEF 2015: Multi-label classification task," in *Working Notes of CLEF 2015 (Cross Language Evaluation Forum)*, Toulouse, France, 2015.

[11] A. Kumar, D. Lyndon, J. Kim, and D. Feng, "Subfigure and Multi-Label Classification using a Fine-Tuned Convolutional Neural Network," in *Working Notes of CLEF 2016 (Cross Language Evaluation Forum)*, Évora, Portugal, 2016.

[12] H. Müller, A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner-Fushman, S. Antani, and I. Eggel, "Overview of the ImageCLEF 2012 Medical Image Retrieval and Classification Tasks," in *Working Notes of CLEF 2012 (Cross Language Evaluation Forum)*, Rome, Italy, 2012.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," in *ACM International Conference on Multimedia*, 2014, pp. 675–678.

[14] A. Krizhevsky, I. Sutskever, and H. Geoffrey E., "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems 25 (NIPS2012)*, pp. 1–9, 2012.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data Mining and Knowledge Discovery Handbook*, 2nd ed., O. Maimon and L. Rokach, Eds. Springer, 2010, ch. 34, pp. 667–685.

[17] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, p. 49, 2010.