Classifying Biomedical Figures by Modality via Multi-Label Learning

Athanasios Lagopoulos, Nikolaos Kapraras, Vasileios Amanatiadis, Anestis Fachantidis, and Grigorios Tsoumakas, *Member, IEEE CS*

***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 introduce novel multi-label modality classification approaches for biomedical figures without segmenting the compound 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 a 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 approach involving compound figure separation into sub-figures. Furthermore, we study how multimodal learning, from both visual and textual features, affects the tasks of classifying biomedical figures by modality and detecting compound figures. Finally, 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 from any device and provide feedback about the modality of a figure classified by the system.**

***Index Terms*— Biomedical images, Image classification, Image retrieval, Modality classification, Multi-label learning, Supervised learning, Text mining**

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 4.7 million articles and is growing at a rapid pace; 489,727 articles were added in 2017. 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, educators and even patients. The *modality* of a figure (e.g. angiography, microscopy), in particular, is a very helpful kind of meta-data for medical retrieval [1]–[3].

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme Human Resources Development, Education and Lifelong Learning in the context of the project Strengthening Human Resources Research Potential via Doctorate Research (MIS-5000432), implemented by the State Scholarships Foundation (IKY).

This research is partly funded by Atypon Systems, LLC.

The authors are with the School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece (e-mail: lathanag, kapraran, vamanati, afa, greg@csd.auth.gr)

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

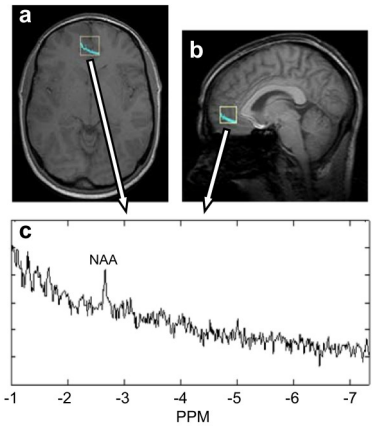

Fig. 1. A compound figure comprising 3 sub-figures: one graph (lower part) and two obtained via magnetic resonance imaging (upper part). Adapted from "A review of imaging techniques for systems biology.", Kherlopian, Armen R., et al. BMC systems biology 2.1 (2008): 74.

About 40% of the figures in PMC are compound, comprising two or more sub-figures in a multi-panel format [4]. Fig. (1) is an example of a compound figure comprising three sub-figures: one graph and two obtained through magnetic resonance imaging. The *figure separation approach* to biomedical figure modality classification first uses a binary model to recognize whether the figure is compound or not. If the figure is not compound, then a single-label 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 [5]–[8]. Then a single-label model is used to predict the modality of each sub-figure. Training and prediction with the figure separation approach are depicted in Fig. (2).

Figure separation is not perfect. Past ImageCLEF benchmarks (2015, 2016), show that 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 single-label 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 [4], [10]–[12].

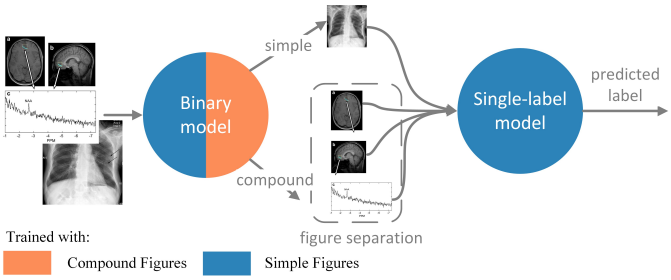

Fig. 2. Training and prediction with the figure separation approach.

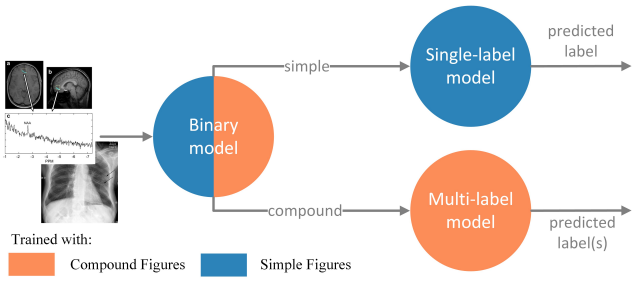

Fig. 3. Training and prediction with the standard multi-label approach.

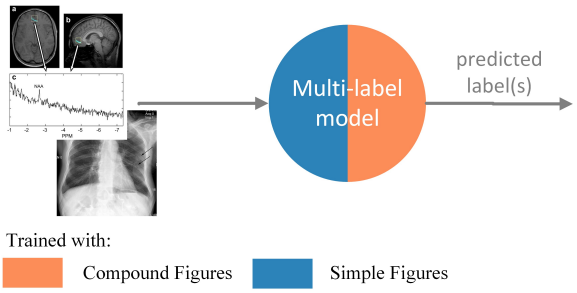

Fig. 4. Training and prediction with the simple multi-label approach.

This article focuses on using multi-label learning for recognizing the modalities that characterize a (potentially compound) biomedical figure without using a figure separation algorithm. We conducted an experimental comparison of the figure separation approach 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) discarding the compound figure detection model from the prediction process. In addition, we examine the performance of visual, textual and multimodal approaches to learning compound figure detection and multi-label classification models. Finally, we describe a publicly available figure retrieval system that runs on top of PMC and incorporates our best modality classification approach.

We envisage the combination of information and knowledge from millions of open-access scientific articles (big data), such as figures and their meta-data, with individual patient data coming from small wearable or disposable sensors (small things) [13], [14] to lead to exciting personalized medical systems and associated business models.

A preliminary version of this work, examining only the performance of visual features and discussing an early version of our system, was presented at the 31st IEEE International Symposium on Computer-Based Medical Systems [15].

The rest of this article is organized as follows. Section (II) presents our multi-label learning approaches and outlines their differences from past work. Section (III) describes the data, the learning algorithms and the evaluation metrics used in our study. Section (IV) experimentally compares the multi-label approaches with the figure separation approach, whereas Section (V) compares the visual, textual and multimodal approaches. Our medical figure retrieval system is described in Section (VI) and conclusions along with future work are presented in Section (VII).

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 figure separation approach in Fig. (2) from using a figure separation module followed by a single-label model to using a multi-label model trained on compound figures. Training and prediction with this *standard multi-label* approach are depicted in Fig. (3).

However, this approach forgoes the use of non-compound, hereafter called *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 simple figures, essentially treating the latter as multi-label examples with just a single label. Training and prediction with this *simple multi-label* approach are depicted in Fig. (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 consider using both compound and simple figures for training the multi-label model in the standard multi-label approach. Training and prediction with this *extended multi-label* approach are depicted in Fig. (5). Only figures predicted by the binary model as compound are passed on to the multi-label model.

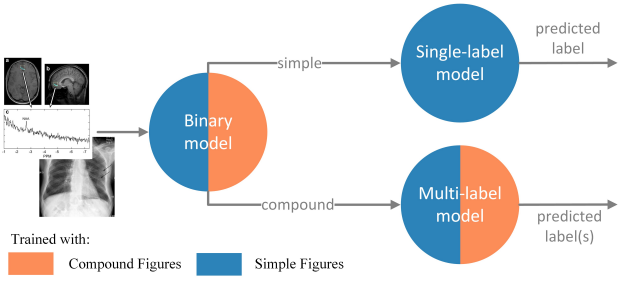

Fig. 5. Training and prediction with the extended multi-label approach.

Algorithm 1 Training steps of the four approaches presented.

Input: X - A set of figures, C - A binary vector specifying if a figure in X is compound (1) or simple (0). $|X| = |C| = n$

Output: M - A set of trained models.

$X_{simple} \leftarrow \{x_i | x_i \in X \text{ if } c_i = 0, c_i \in C\}, i = [1, \ldots, n]$
$X_{compound} \leftarrow \{x_i | x_i \in X \text{ if } c_i = 1, c_i \in C\}$
if approach == "figure separation" **then**
    binary $\leftarrow trainBinary(X)$
    single $\leftarrow trainSingle(X_{simple})$
    return binary, single
else if approach == "standard multi-label" **then**
    binary $\leftarrow trainBinary(X)$
    single $\leftarrow trainSingle(X_{simple})$
    multi $\leftarrow trainMulti(x_{compound})$
    return binary, single, multi
else if approach == "simple multi-label" **then**
    multi $\leftarrow trainMulti(x_{compound})$
    return multi
else if approach == "extended multi-label" **then**
    binary $\leftarrow trainBinary(X)$
    single $\leftarrow trainSingle(X_{simple})$
    multi $\leftarrow trainMulti(X)$
    return binary, single, multi

Algorithm 2 Prediction steps of the four approaches presented.

Input: X - Set of unclassified figures, $binary$ - the binary model, $single$ - the single-label model, $multi$ - the multi-label model. Each model is trained with respect to the approach as shown in Algorithm 1

Output: L: Set of binary vectors, one for each figure in X, with the predicted modalities of the unclassified figures X.

if approach == "simple multi-label" **then**
    $L \leftarrow multi(X)$
else
    $C \leftarrow binary(X)$
    $X_{simple} \leftarrow \{x_i | x_i \in X \text{ if } c_i = 0, c_i \in C\}, i = [1, \ldots, n]$
    $X_{compound} \leftarrow \{x_i | x_i \in X \text{ if } c_i = 1, c_i \in C\}$
    if approach == "figure separation" **then**
        $X_{subfigures} \leftarrow segmentation(X_{compound})$
        $L \leftarrow single(X_{simple} + X_{subfigures})$
    else if approach == "standard multi-label" or "extended multi-label" **then**
        $L_{simple} \leftarrow single(X_{simple})$
        $L_{compound} \leftarrow multi(X_{compound})$
        $L = L_{simple} + L_{compound}$
return L

Algorithms (1) and (2) describe, respectively, the training and prediction steps of all the approaches presented.

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 [4], [11]. 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 [4], [10]–[12] focused solely on multi-label classification of compound figures and did not look at the alternative architectures we discuss here, which take into consideration simple figures too, in order to address the full real-world problem of classifying figures (whether they are compound or not) 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 single-label model per sub-figure. None looked into exploiting both visual and textual features.

III. Main Aspects of our Experimental Studies

This section initially describes the data we used for our experimental studies and outlines the features extracted. It then details the base learning algorithms underlying our approaches and the evaluation metrics we used to compare them.

A. Original 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, along with their captions and corresponding PMC article IDs, of which 12,338 (59%) are compound and 8,647 (41%) simple. 1,568 (13%) of the compound figures are annotated with one or more of the 30 classes of the hierarchy in Fig. (6), which refer to biomedical illustrations and types of diagnostic images (radiology, visible light photography, printed signals, microscopy, 3D reconstructions). Furthermore, the 6,776 sub-figures of these 1,568 compound figures are also given along with their annotation with one of the 30 classes of the hierarchy in Fig. (6).

B. Feature Extraction

We extract visual features from the JPEG file of each figure and sub-figure using the Caffe framework [16]. We used the BVLC CaffeNet deep learning model[3], which is a replication of the model described in [17] and has been trained with 1.2 million high-resolution images. For each JPEG file, 4,096 features were extracted via the fc7 (inner product or fully connected) level.

We extract textual features from a figure's caption and/or the sentence, within the full-text of the figure's article, containing a reference to the figure. In specific, we compute term frequencies of word n-grams (unigrams and bigrams) using the TFIDFVectorizer class of scikit-learn[4]. Inverse document-frequency reweighing was disabled and English stop-word removal was used. The rest of the parameters were set to default values.

Note that we do not extract textual features from sub-figures since their caption is not provided. The caption of a sub-figure can be considered as a text snippet (sub-caption) of

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

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

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

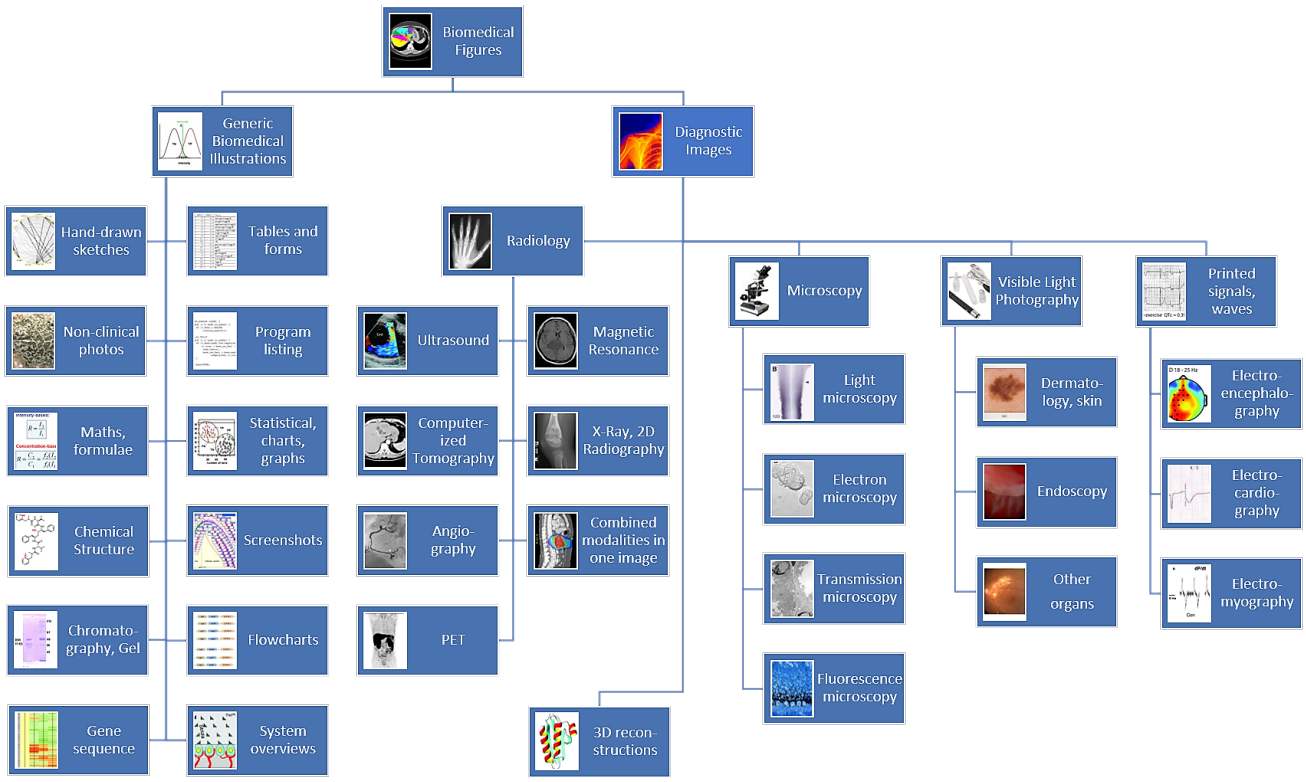

Fig. 6. Hierarchy of 30 modality classes concerning various types of diagnostic images and generic biomedical illustrations.

the caption of the compound figure from which the sub-figure was extracted. However, caption segmentation is out of the scope of this work.

C. Learning Algorithms

We employ logistic regression to learn all models with visual features and linear support vector machines (SVMs) for the models with textual features. We use the scikit-learn library and default parameter settings for both algorithms (SVM: cost parameter equal to 1 - squared hinge loss function - L2 penalization, Logistic Regression: cost parameter equal to 1 - tolerance of $1e-4$ - $L1$ penalization). The One-vs-Rest (OvR) transformation was used to decompose the single-label learning problem into multiple binary classification tasks, and similarly, the binary relevance (BR) transformation was used to decompose the multi-label learning problem [18].

D. Evaluation Metrics

We use micro- and macro-averaging to compute binary evaluation metrics, such as recall, precision and F-measure, across all labels in the single-label and multi-label tasks. Micro-averaging calculates metrics globally by counting all true positives, false negatives and false positives, whereas macro-averaging calculates metrics per class/label and then takes the mean across all classes/labels. In a single-label setting, the number of false positives equals the number of false negatives making micro-averaging recall, precision and F-measure metrics equivalent. In the multi-label task, we further use samples-averaging, which calculates metrics per instance and then takes the mean across all instances. In the compound figure detection task, we use balanced accuracy and g-mean along with per class recall, precision and F-measure to avoid optimistic estimations due to imbalance. For all the metrics above, we use the approach described in [19] to avoid biased cross-validated results. All approaches and models were evaluated using 10-fold cross-validation. The above binary evaluation measures are calculated, based on the values of the confusion matrix: true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn), as follows:

$$\text{Recall} = \frac{tp}{tp+fn} \qquad \text{Precision} = \frac{tp}{tp+fp}$$

,

$$\text{F-measure} = \frac{2*tp}{2*tp+fn+fp}$$

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{tp}{tp+fn}+\frac{tn}{tn+fp}\right)$$

$$\text{G-mean} = \frac{tp}{\sqrt{(tp+fp)(tp+fn)}}$$

Let $tp_\lambda, fp_\lambda, tn_\lambda$ and fn_λ be the number of true positives, false positives, true negatives and false negatives after binary evaluation for a label λ. The macro-averaged and micro-averaged versions of a binary evaluation measure B are calculated as follows:

$$B_{\text{macro}} = \frac{1}{q}\sum_{\lambda=1}^{q} B\left(tp_\lambda, fp_\lambda, tn_\lambda, fn_\lambda\right)$$

$$B_{\text{micro}} = B\left(\sum_{\lambda=1}^{q} tp_\lambda, \sum_{\lambda=1}^{q} fp_\lambda, \sum_{\lambda=1}^{q} tn_\lambda, \sum_{\lambda=1}^{q} fn_\lambda\right)$$

IV. Experimental Comparison of Multi-Label Modality Classification Approaches

A. Data

As we mentioned in Section I, about 40% of the available figures in PMC are compound. However, the medical task of ImageCLEF 2016 did not provide annotations for the 8,647 simple figures it delivered. Therefore, in order to simulate training and test sets following the distribution of PMC, we adopt the following process. We first split the 1,568 compound figures into 10 equally sized disjoint subsets (folds). At each fold, we keep 40% of the figures as they are and replace the rest with their respective sub-figures, which assume the role of simple figures. This process further ensures that sub-figures of the same figure stay within the same fold in order to avoid information leakage from a training set to a test set. Since we can't extract textual features from sub-figures, we only use visual features for comparing the different modality classification approaches.

B. Results

Table (I) shows the micro-, macro- and samples-averaged F-measure for the standard, simple and extended multi-label approaches, for an approach assuming perfect figure separation into sub-figures, and for a recent figure separation approach with state-of-the-art results in segmenting biomedical figures of academic publications based on data from ImageCLEF 2013, 2015 and 2016 [8].

We first notice that the simple multi-label approach is the worst of all our proposed approaches, highlighting the importance of having a compound figure detection model. The standard and extended multi-label approaches have similar micro-averaged F-measure, but the extended one leads to 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 equivalently, whereas 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 improvement in the standard and extended multi-label approaches. We did not find consistent improvements across classes, as for 7 (14) classes the F-measure was better with the standard (extended) multi-label approach, and for 9 classes the F-measure did not change between approaches. In 7 out of the latter 9 classes, the F-measure was actually zero and this was due to the very small number (less than 5) of training examples. Table (II) shows the 5 classes with the highest improvement (upper part) and deterioration (lower part) of their F-measure when switching from the standard to the extended multi-label approach.

We then looked at the Pearson correlation between the F-measure of the standard approach and the percentage of change in F-measure between the two approaches. We found a moderate correlation of ($r \approx -0.41$). 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. If we only consider the 10 classes of Table (II), then this correlation rises to approximately -0.51.

We also notice that the perfect figure separation approach achieves the best results in all measures, while the real-world figure separation approach achieves the worst micro- & samples-average F-measure. The segmentation part of the latter approach achieves an accuracy of 81.78%, computed as described in [9]. We project linearly that a segmentation accuracy of $\approx$ 96.7% is needed for a figure separation approach to achieve the same samples-average F-measure as the extended multi-label approach. Considering the above and that the differences between the perfect figure separation and the extended multi-label approach are quite small, we conclude that the extended multi-label approach is very promising. 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 [18].

Table (III) shows per class recall, precision and F-measure of the binary compound figure detection model, common in the *figure separation*, *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

TABLE I
Results comparing the four approaches

	F-measure		
	Macro	Micro	Samples
Perfect figure separation	**0.3631**	**0.7875**	**0.7979**
Figure separation	0.3194	0.7646	0.7539
Standard multi-label	0.3278	0.7789	0.7828
Simple multi-label	0.3077	0.7733	0.7268
Extended multi-label	0.3426	0.7785	0.7901

TABLE II
Top-5 improved and deteriorated classes when switching from standard (Std) to extended (Ext) multi-label approach

	F-measure		Diff %
	Std	Ext	
Flowchart	0.0333	0.0833	150.15
Endoscopy	0.2167	0.2833	30.73
Combined Modalities	0.1900	0.2200	15.78
Non-Clinical	0.3447	0.3935	14.15
Other Organs	0.1509	0.1692	12.12
Ultrasound	0.2952	0.2452	-16.93
Chemical Structure	0.3689	0.3409	-7.59
Screenshot	0.1667	0.1619	-2.87
Transmission Microscopy	0.5670	0.5565	-1.85
Light Microscopy	0.8030	0.7968	-0.77

TABLE III
RESULTS FOR THE COMPOUND DETECTION MODEL

	Compound	Simple
Recall	0.7905	0.9803
Precision	0.8601	0.9683
F-measure	0.8238	0.9742
Balanced Accuracy	0.8854	
G-mean	0.9448	

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 single-label model common in all approaches, apart from the simple multi-label one, as measured in the pipeline of the standard/extended multi-label approaches. Evaluation is based on the sub-figures of the data for this section's experiments, which as explained at the beginning of this section come from the 60% of all compound figures. Note that in single-label tasks, false positives equal false negatives and so precision equals recall and F-measure.

V. EXPERIMENTAL COMPARISON OF VISUAL, TEXTUAL AND MULTIMODAL APPROACHES

A. Data

We experiment with two constituent models of the modality classification approaches that do not involve sub-figures and therefore textual features can be readily extracted: i) the compound detection model (binary model), and ii) the multi-label classification model of the standard multi-label approach. We use the dataset of 20,985 figures (compound, simple) for the first model and the 1,568 compound figures for the second model.

B. Multimodal Approaches

We experiment with both early and late fusion of the textual and visual modalities of our data. We adopt the standard early fusion approach which constitutes in merging the visual and textual features. For late fusion, we adopt a stacking approach [20]. In general, stacking trains a meta-level classifier using as input the decision function scores of a number of base level classifiers. In multimodal learning, each of the base classifiers is trained on a different data modality. Stacking is also used as a more advanced, compared to BR, multi-label learning method aiming to exploit dependencies between the multiple labels [21]. In this case, each of the base classifiers corresponds to a different label, while at the meta-level one classifier per label is trained for a second time. BR is applied twice in this typical multi-label stacking framework.

TABLE IV
RESULTS FOR THE SINGLE-LABEL MODEL

	Macro	Micro
Recall	0.3811	0.8044
Precision	0.4412	0.8044
F-measure	0.4090	0.8044

For the compound figure detection model, we first train two base binary classifiers, one using only the visual features and one using only the textual features. Then, a meta binary classifier is trained using as inputs the scores of these two classifiers. For the multi-label classification model, we first train 60 binary classifiers, one for each class and data modality pair, i.e. 30 using visual and 30 using textual features. Then, we learn 30 meta level classifiers, one for each class, taking as input the 60 scores of all base-level classifiers. We, therefore, exploit stacking simultaneously as both a multi-label and a multimodal learning approach. For a fair comparison against single modal learning, we also employ stacking as a multi-label approach using only visual and only textual features. We use a linear support vector machine algorithm for the meta-classifiers.

C. Results

Table (V) shows the balanced accuracy, G-mean and average F-measure for the compound figure detection model comparing different setups of the textual features. We test the different setups by using a validation set comprising of the 20% of the 20,985 figures dataset. We first notice that the best results are achieved, in all the measures, when using both the caption and the article's content in combination with the use of both unigrams and bigrams. Also, the fact that the use of the content in tandem with the caption achieves the best scores, regardless the n-gram representation, shows that the content of the article holds information that can help a model better discriminate compound from simple figures.

Table (VI) compares the single modal and multimodal approaches we have discussed, using the best textual features setup (caption, content, unigram, and bigram). We see that the visual model is clearly better than the textual one in both balanced accuracy and average F-measure. In addition, the results show that multimodal learning can achieve higher scores in detecting compound figures. Especially the late fusion model increases the average F-measure by 3.76% (7.84%) compared to the visual (textual) model.

Table (VII) shows the macro-, micro-, and samples-averaged F-measure for the multi-label classification model comparing different setups of textual features. We test the different setups by using a validation set comprising of the 20% of the 1,568 compound figures dataset. Here, the best model uses just the caption of the figure with both unigram and bigram representations (micro- and samples-averaging) and just unigram (macro-averaging). We attribute this finding to the short length of the content's text, just one sentence, which may be inadequate to produce meaningful features for all the labels of a figure. We further notice that the absence of the caption severely affects prediction accuracy.

Table (VIII) compares single modal (with and without stacking) learning with multi-modal learning, using the best setup for the textual data modality. It is again observed that the visual model achieves better results than the textual one. However, the stacking technique improves the textual model

TABLE V
RESULTS FOR THE COMPOUND FIGURE DETECTION MODEL COMPARING DIFFERENT SETUPS OF TEXTUAL FEATURES, SORTED BY AVERAGE F-MEASURE IN DESCENDING ORDER.

Text Source		Representation		Balanced Accuracy	G-mean	Average F-measure
Caption	Content	Unigram	Bigram			
✓	✓	✓	✓	**0.7202**	**0.7269**	**0.7312**
✓	✓	✓		0.7180	0.7023	0.7228
✓	✓		✓	0.7084	0.7007	0.7134
✓		✓		0.6973	0.6942	0.7025
✓		✓	✓	0.6862	0.7179	0.6993
	✓	✓	✓	0.6950	0.6874	0.6984
	✓	✓		0.6904	0.6680	0.6884
✓			✓	0.6642	0.7027	0.6753
	✓		✓	0.6755	0.6373	0.6662

TABLE VI
RESULTS OF THE COMPOUND DETECTION MODEL COMPARING SINGLE-MODAL TO MULTIMODAL LEARNING.

	Balanced Accuracy	G-mean	Average F-measure
Visual	0.8255	0.7273	0.8255
Textual	0.7592	0.7174	0.7942
Early Fusion	0.8138	0.7471	0.8285
Late Fusion	**0.8280**	**0.7882**	**0.8565**

TABLE VII
RESULTS OF THE MULTI-LABEL MODEL COMPARING DIFFERENT SETUPS OF TEXTUAL FEATURES, SORTED BY SAMPLES F-MEASURE IN DESCENDING ORDER.

Text Source		Representation		F-measure		
Caption	Content	Unigram	Bigram	Macro	Micro	Samples
✓		✓	✓	0.1888	**0.6584**	**0.6377**
✓	✓	✓	✓	0.1603	0.6545	0.6335
✓	✓	✓		0.1571	0.6458	0.6244
✓		✓		**0.2156**	0.6461	0.6193
	✓	✓	✓	0.1291	0.5854	0.5724
	✓	✓		0.1299	0.5859	0.5634
✓			✓	0.1412	0.5787	0.5615
✓	✓		✓	0.0973	0.5307	0.5176
	✓		✓	0.0685	0.4800	0.4759

TABLE VIII
RESULTS OF THE MULTI-LABEL MODEL COMPARING SINGLE-MODAL AND MULTIMODAL LEARNING.

	F-measure		
	Macro	Micro	Samples
Visual	**0.2860**	0.6906	0.6999
Textual	0.1779	0.6521	0.6295
Visual (Stacking)	0.2621	0.6670	0.6777
Textual (Stacking)	0.2095	0.6788	0.6761
Early Fusion	0.2282	0.6672	0.6718
Late Fusion	0.2439	**0.7115**	**0.7341**

but considerably harms the visual one. We believe that this irregularity which occurs when using stacking is due to the rarity of some modality classes. Rare classes, in contradiction to frequent ones, create weak correlations with other classes whereas stacking techniques exploit dependencies between classes. Thus, when stacking, models lose their ability to classify rare classes and only boost their ability in classifying frequent classes. Visual models can discriminate rare classes much better than the textual (higher macro-average, similar micro-average), but their ability is harmed when using stacking. This also explains the decrease of macro-averaging F-measure (-8.36%) and a limited decrease of micro-averaging F-measure (-3.54%) when comparing the visual model with and without stacking. This behavior is also noticed in the case of the late fusion model, which shows a significant increase in micro- and samples- averaged F-measure in comparison to all the other models, but also a decrease in macro-averaged F-measure compared to the visual model. Overall, multimodal learning increases the accuracy of the multi-label classification model but the underlying stacking technique harms the discrimination capabilities in rare classes. A hybrid technique using late fusion only for the frequent classes and visual for the rare ones is expected to achieve even better results.

VI. THE MEDIEVAL SYSTEM

We embedded our approaches in a Web application 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 and textual features and then classifies the modality using the best of the proposed pipelines. Specifically, it first uses a compound detection model trained on the 20,985 figures data set to identify if the image is compound or simple. Then, if the figure is compound a multi-label model is invoked which is trained with the full set of 1,568 figures. Both models are trained using the late fusion multimodal method. Finally, if the figure is classified as simple, a single-label model is invoked which is trained on the 6,776 sub-figures. Since the training involves sub-figures, this model only employs visual features. The modality predictions (ground truth for the training set) along with the figure's caption, unique PMC ID, 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.

Users can give feedback about the modalities of a figure appearing in search results. They can add or remove any of the modalities of a figure and submit their changes for review

[5]http://intelligence.csd.auth.gr/medieval

[6]https://angularjs.org/

[7]https://www.ncbi.nlm.nih.gov/pmc/tools/oai/

[8]http://lucene.apache.org/solr/

by an expert. The motivation behind this function of the system is to enable the crowdsourcing of more annotations, with the ultimate goal of improving the accuracy of our classification approaches. Not only will the fresh data allow us to use actual simple figures for our models, and thus make use of textual features, but it will also increase the discriminative capability of our models, particularly for rare modality classes. Toward this, we added gamification components such as weekly/monthly leader-boards and achievement batches to attract more users and experts on giving their feedback [22].

VII. 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 compound and simple 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 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 by building separate models to classify the same underlying modalities. The experimental 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 processes. The *extended multi-label* approach showed particularly promising results, only slightly worse than using a perfect figure separation approach. Furthermore, the comparison of single-modal and multi-modal learning shows that the addition of textual features can greatly improve the classification scores. 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 and give feedback on the modalities of a classified figure.

In the future, we plan to investigate how we can further improve our approaches by using more advanced multi-label learning techniques, such as ensembles of classifier chains. In addition, we aim to look further into the available pre-trained deep learning models for extracting better fitted visual features. Furthermore, we will consider the use of weak supervision techniques [23] and specifically how we can combine noisy supervision and co-training. Our future plans also include a further extension of MEDIEVAL so as to make use of the users' feedback by employing active learning techniques. The application will be able to explicitly request the feedback that will mostly benefit the classification system.

Acknowledgment

We would like to thank the anonymous reviewers for their constructive comments and valuable suggestions.

References

[1] 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.

[2] 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.

[3] 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.

[4] A. García Seco de Herrera, H. Müller, and S. Bromuri, "Overview of the ImageCLEF 2015 medical classification task," in *Working Notes of CLEF 2015*, Toulouse, France, 2015.

[5] 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.

[6] 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.

[7] 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.

[8] P. Li, X. Jiang, C. Kambhamettu, and H. Shatkay, "Compound image segmentation of published biomedical figures," *Bioinformatics*, vol. 34, no. 7, pp. 1192–1199, 2017.

[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*, Toulouse, France, 2015.

[11] 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*, Évora, Portugal, 2016.

[12] 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*, Évora, Portugal, 2016.

[13] S. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K.-S. Kwak, "The internet of things for health care: a comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015.

[14] M. Hassanalieragh, A. Page, T. Soyata, G. Sharma, M. Aktas, G. Mateos, B. Kantarci, and S. Andreescu, "Health monitoring and management using internet-of-things (iot) sensing with cloud-based processing: Opportunities and challenges," in *2015 IEEE International Conference on Services Computing*. IEEE, 2015, pp. 285–292.

[15] A. Lagopoulos, A. Fachantidis, and G. Tsoumakas, "Multi-label Modality Classification for Figures in Biomedical Literature," *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, vol. 2017-June, no. i, pp. 79–84, 2017.

[16] 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.

[17] 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.

[18] 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.

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

[20] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, feb 1992.

[21] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas, "Multi-target regression via input space expansion: treating targets as inputs," *Machine Learning*, vol. 104, no. 1, pp. 55–98, 2016.

[22] B. Morschheuser, J. Hamari, J. Koivisto, and A. Maedche, "Gamified crowdsourcing: Conceptualization, literature review, and future agenda," *International Journal of Human Computer Studies*, vol. 106, pp. 26–43, oct 2017.

[23] J. Hernández-González, I. Inza, and J. A. Lozano, "Weak supervision and other non-standard classification problems: A taxonomy," *Pattern Recognition Letters*, vol. 69, pp. 49–55, 2016.