# Overview of the ImageCLEF 2016 Medical Task

Alba G. Seco de Herrera[1], Roger Schaer[2], Stefano Bromuri[3], and Henning Müller[2],

[1] Lister Hill National Center for Biomedical Communications, National Library of Medicine,Bethesda, USA;
[2] University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland;
[3] Open University of the Netherlands, The Netherlands
albagarcia@nih.gov

**Abstract.** ImageCLEF is the image retrieval task of the Conference and Labs of the Evaluation Forum (CLEF). ImageCLEF has historically focused on the multimodal and language–independent retrieval of images. Many tasks are related to image classification and the annotation of image data as well. The medical task has focused more on image retrieval in the beginning and then retrieval and classification tasks in subsequent years. In 2016 a main focus was the creation of meta data for a collection of medical images taken from articles of the the biomedical scientific literature. In total 8 teams participated in the four tasks and 69 runs were submitted. No team participated in the caption prediction task, a totally new task.
Deep learning has now been used for several of the ImageCLEF tasks and by many of the participants obtaining very good results. A majority of runs was submitting using deep learning and this follows general trends in machine learning. In several of the tasks multimodal approaches clearly led to best results.

**Keywords:** ImageCLEFmed, compound figure detection, multi–label classification, figure separation, modality classification, caption detection

## 1 Introduction

ImageCLEF has organized image retrieval evaluation campaigns since 2003 and a medical task was added in 2004 [1, 2]. With a focus on multimodal and language–independent retrieval of images, the databases used have evolved strongly over the years and many of the datasets have gotten larger as well. Several medical tasks have been organized over the years [3–5], ranging from the classification of medical images to retrieval of single images or entire cases. This year's tasks are an evolution from the tasks that were organized in 2015 [6]. The main objective with the current data sets and tasks is to make the large amount of visual content that is shared in the biomedical open access literature available in an easier way by generating meta data. PubMed Central[4] (PMC) makes a large

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

amount of currently over 4 million articles available including text and figures in a structured form. The collection is growing strongly with over 200'000 articles being added in 2014 alone and with a quickly increasing tendency. Basically no metadata are available for the figures besides the figure captions and global information on the articles including global MeSH (Medical Subject Headings) terms. A major problem is that about half of the available figures contain more than one subfigure, so are compound or multi–pane figures. The tasks in 2016 aims at first detecting, whether a figure is a compound figure, then trying to separate the compound figures into their parts or extract image type information for all subfigures of a compound figure. Then, a modality classification tries to detect the image type, that ranges from medical modalities (e.g. X–ray, MRI, CT) to general image types such as graphs and flow charts. All these tasks can help to generate metadata for the almost 4 million images available via PMC. Including the extracted subfigures this will likely amount to over 10 million medical figures that are available and currently only little exploited.

This article first describes the five tasks that were organized in 2016, then describes the data sets, ground truth and participation. The conclusions summarize the main lessons learned from the evaluation campaign. Finally, a little outlook is given into the limitations and a possible future of the task.

## 2 Tasks, Data Sets, Ground Truth, Participation

### 2.1 The Tasks in 2016

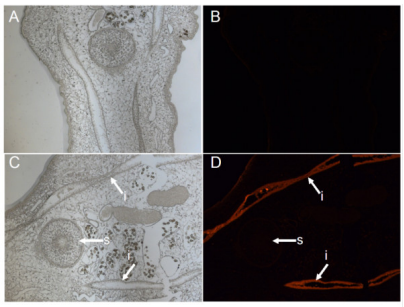
Five subtasks were organized in 2016:

- compound figure detection;
- compound figure separation;
- multi–label classification with image types;
- subfigure classification into image types;
- caption prediction.

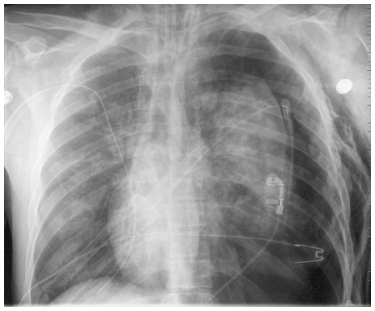
This section gives an overview of each of the five subtasks.

**Compound Figure Detection** As a first step for the retrieval of compound figures and its subfigures, compound figure detection is necessary. This subtask was introduced in 2015 and the goal is to identify whether a figure is a compound figure or not (see Figure 1). The task is not easy, as compound figures can or not have dominant subfigures and do not always have clear separating lines. The subtask provides a set of compound and non–compound figures from the biomedical literature of PMC.

**Compound Figure Separation** The goal of this subtask is to separate the compound figures into subfigures to be able to work with the subfigures independently. This subtask was introduced in 2013. Figure 2 shows a simple example of a compound figure separated by blue lines. There are many more challenging examples where the separating lines are not straight or where a large number of sub figures is put into a single figure.




(a) Compound figure.



(b) Non–compound figure.

**Fig. 1.** Examples of compound and non–compound figures.

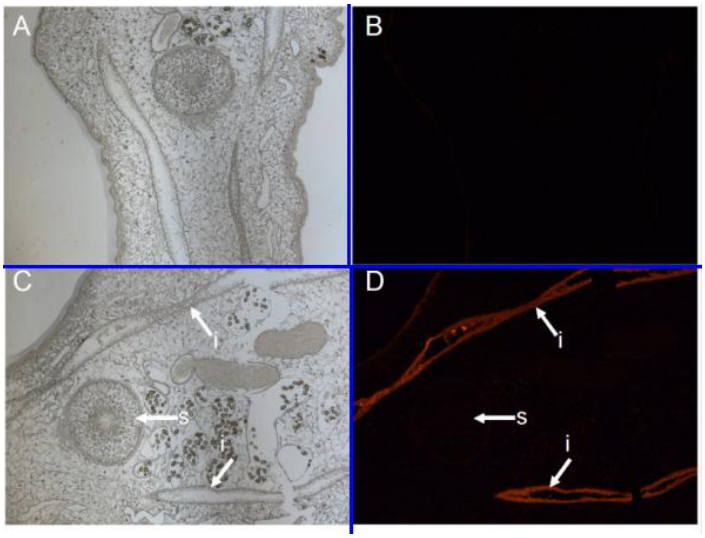



**Fig. 2.** Example of a compound figure separated into subfigures, showing the blue separation lines.

**Multi–label Classification** Compound figures are often an aggregate belonging to multiple classes, as they can show multiple perspectives concerning a medical problem. This aggregation is not random, a compound figure is an aggregation of sub–figures representing a relationship with a clear semantic meaning. Goal of the task was to see whether it is possible to learn the components of a figure without learning from single images representing the image types, but from other labelled compound figures. Techniques such as deep learning should work well on such tasks.

Figure 3 is an example of such an image in which multiple perspectives of the same set of cells are shown.

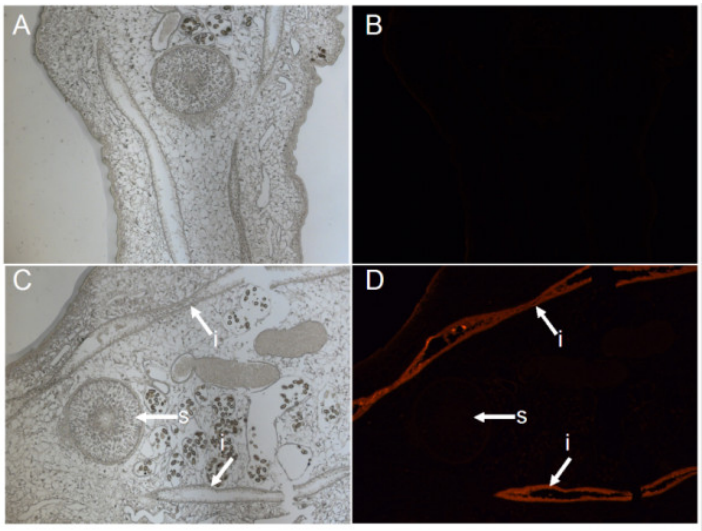



**Fig. 3.** Example of a compound figure with multiple related subparts that was used for the multi–label classification subtask. This figure contain the labels "Light microscopy" (subfigures A and C) and "Fluorescence microscopy" (subfigures B and D).

Bromuri et al. [7] formulates the general multi–label problem as follows:

Let $X$ be the domain of observations and let $L$ be the finite set of labels. Given a training set $T = \{(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)\}$ $(x_i \in X, Y_i \subseteq L)$ i.i.d. drawn from an unknown distribution $D$, the goal is to learn a multi-label classifier $h : X \to 2^L$. However, it is often more convenient to learn a real–valued scoring function of the form $f : X \times L \to \mathbb{R}$. Given an instance $x_i$ and its associated label set $Y_i$, a working system will attempt to produce larger values for labels in $Y_i$ than those that are not in $Y_i$, i.e. $f(x_i, y_1) > f(x_i, y_2)$ for any $y_1 \in Y_i$ and $y_2 \notin Y_i$. By the use of the function $f(\cdot, \cdot)$, we can obtain a multi–label classifier: $h(x_i) = \{y | f(x_i, y) > \delta, y \in L\}$, where $\delta$ is a threshold to infer from the training set. The function $f(\cdot, \cdot)$ can also be adapted to a ranking function $rank_f(\cdot, \cdot)$,

which maps the outputs of $f(x_i, y)$ for any $y \in L$ to $\{1, 2, ..., |L|\}$ such that if $f(x_i, y_1) > f(x_i, y_2)$ then $rank_f(x_i, y_1) < rank_f(x_i, y_2)$.

Multi–label performance measures are generally different from those used in the single label tasks. In [6], we introduced the Hamming loss. The Hamming loss evaluates how many times an observation–label pair is misclassified. The score lies between 0 and 1, where 0 is the best:

$$hloss_S(h) = \frac{1}{m} \sum_{i=1}^{m} \frac{|h(x_i) \triangle Y_i|}{|L|}. \quad (1)$$

$\triangle$ represents the symmetric difference.

In 2016, we also introduce the mean of the F–Measures of all the labels belonging to a figure, where for each $i$ label we define the F–Measure to be:

$$F\text{–}Measure_i = 2 * \frac{precision_i * recall_i}{precision_i + recall_i} \quad (2)$$

This allows us to understand if there is an unbalanced distribution of the labels that leads to the classifier overfitting to the majority class.

**Subfigure Classification** The figure classification task was already run in a slightly different configuration from 2011 to 2013. In 2015 and 2016 the subtask focuses on the modality classification of subfigures extracted from the compound figures distributed for the multi–label classification subtask. This subtask aims to classify figures into the 30 classes of the hierarchy shown in Figure 4. The class codes with descriptions are the following ([Class code] Description):

- $[Dxxx]$ Diagnostic images:
  - $[DRxx]$ Radiology (7 categories):
  - $[DRUS]$ Ultrasound
  - $[DRMR]$ Magnetic Resonance
  - $[DRCT]$ Computerized Tomography
  - $[DRXR]$ X–Ray, 2D Radiography
  - $[DRAN]$ Angiography
  - $[DRPE]$ PET
  - $[DRCO]$ Combined modalities in one image
- $[DVxx]$ Visible light photography (3 categories):
  - $[DVDM]$ Dermatology, skin
  - $[DVEN]$ Endoscopy
  - $[DVOR]$ Other organs
- $[DSxx]$ Printed signals, waves (3 categories):
  - $[DSEE]$ Electroencephalography
  - $[DSEC]$ Electrocardiography
  - $[DSEM]$ Electromyography
- $[DMxx]$ Microscopy (4 categories):
  - $[DMLI]$ Light microscopy
  - $[DMEL]$ Electron microscopy

  - [$DMTR$] Transmission microscopy
  - [$DMFL$] Fluorescence microscopy
- [$D3DR$] 3D reconstructions (1 category)
- [$Gxxx$] Generic biomedical illustrations (12 categories):
  - [$GTAB$] Tables and forms
  - [$GPLI$] Program listing
  - [$GFIG$] Statistical figures, graphs, charts
  - [$GSCR$] Screenshots
  - [$GFLO$] Flowcharts
  - [$GSYS$] System overviews
  - [$GGEN$] Gene sequence
  - [$GGEL$] Chromatography, Gel
  - [$GCHE$] Chemical structure
  - [$GMAT$] Mathematics, formula
  - [$GNCP$] Non–clinical photos
  - [$GHDR$] Hand–drawn sketches

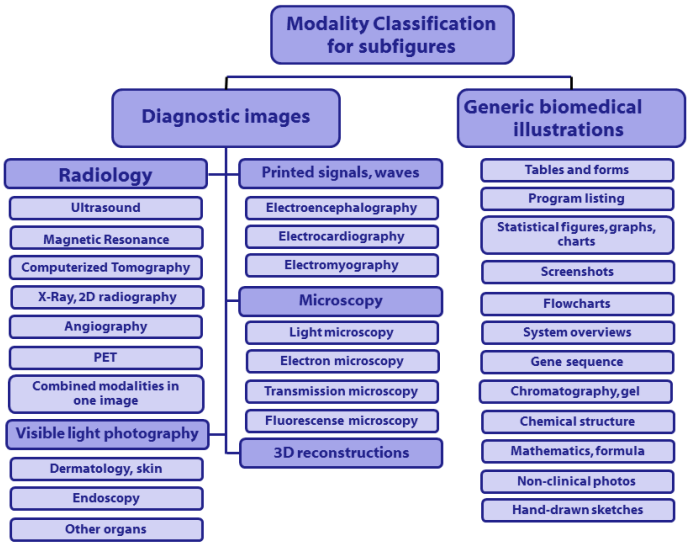
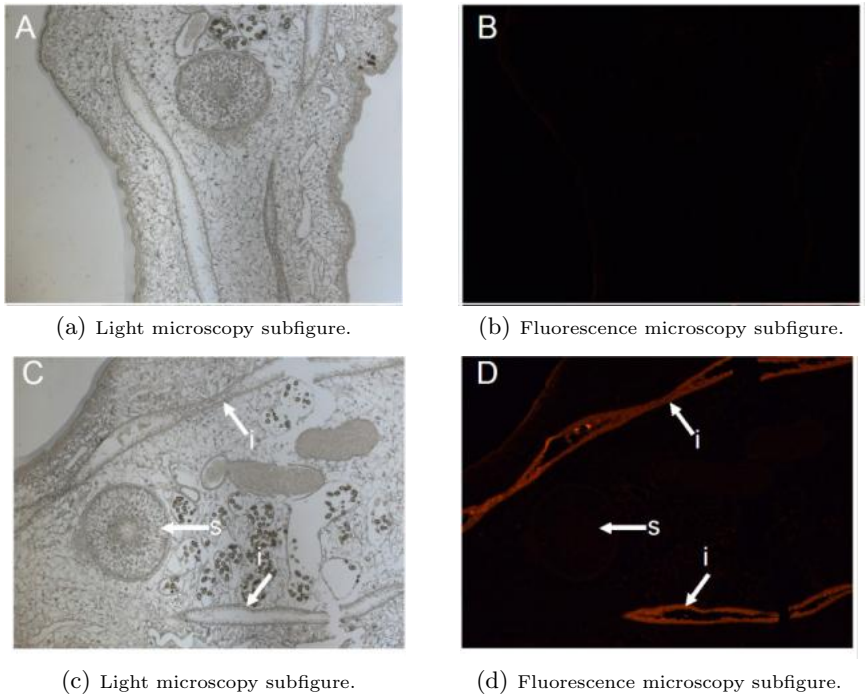



**Fig. 4.** The image class hierarchy that was developed for document images occurring in the biomedical open access literature [8].

Figure 5 shows four subfigures from a compound figured from two different classes.




(a) Light microscopy subfigure. (b) Fluorescence microscopy subfigure.

(c) Light microscopy subfigure. (d) Fluorescence microscopy subfigure.

**Fig. 5.** Example of subfigures belonging to a compound figure.

**Caption Prediction** Thanks to the technical advances of cloud computing many large and data intensive applications have become possible. Modern GPUs (Graphical Processing Units) have made massively parallel computing of simple operations possible and lead to a revival of methods based on neural networks with more complex and deeper architectures. There has been a strong hype around such Deep Learning techniques [9]. One of the most successful application of Deep Learning is that of deep captioning [10].

The purpose of the caption prediction task is to mimic the ability of a medical professional to recognize figures in a medical text and provide a description of these figures. We believe that this is be an important task that can lead to future applications in medical image information retrieval.

### 2.2 Datasets

The dataset used in this task is a subset of images contained in articles from the biomedical literature extracted from the PMC. The trainining sets were obtained merging the training and test sets of the ImageCLEFmed 2015 subtasks [6]. Therefore, in 2016 a larger number of figures were distributed than in 2015. Image captions were also provided in addition to all images. For the compound figure detection subtask 21,000 figures were labelled as compound

figures or non–compound for the training set and 3,456 for the test set. A subset of the compound figures of the compound figure detection subtask was distributed to be separated into subfigures for the figure separation subtask. 6,783 and 1,614 were distributes as training and test sets, respectively. In 2016, more stitched compound figures were added making the subtask more complicated. For the multi–label classification, a subset of the compound figures were distributed containing 1,568 in the training set and 1,083 in the test set. These compound figures were separated into subfigures and distributed for the subfigure classification subtask. The naming of the subfigures was done in a way that if the compound figure ID is "1297-9686-42-10-3", then the corresponding subfigure IDs are "1297-9686-42-10-3-1", "1297-9686-42-10-3-2", "1297-9686-42-10-3-3" and "1297-9686-42-10-3-4" on the case of four subfigures. This resulted in 6,776 subfigures in the training set and 4,166 subfigures in the test set.

The data distributed to the participants for the caption prediction subtask involved 10,000 images from diagnostic imaging category and relative captions. We figured that diagnostic images might be of the highest relevance in this context. The test set comprised another 10,000 diagnostic images but the captions were not included for these images.

### 2.3 Participation

72 groups registered and obtained access to the data. The same number of groups as in 2015 submitted results to the medical task (8 groups in total). Groups participated from four continents, so the regional spread was high.

Despite that the number of groups that registered in 2016 being smaller than in 2015, the number of submitted runs increased. 15 runs were submitted to the compound figure detection task, 3 runs to the multi–label classification task, 9 runs to the figure separation task and 42 runs to the subfigure separation task. There were unfortunately no participants in the new caption prediction task.

The following groups submitted at least one run:

- BMET (Institute of Biomedical Engineering and Technology, University of Sydney, Australia);
- CIS UDEL (Computer & Information Sciences, University of Delaware, Newark, USA);
- DUTIR (Department of Computer Science and Engineering, Dalian University of Technology, China.)*;
- FHDO BCSG (FHDO Biomedical Computer Science Group, University of Applied Science and Arts, Dortmund, Germany);
- IPL (Athens University of Economics and Business, Greece);
- MLKD (Department of Informatics, Aristotle University of Thessaloniki, Greece)*;
- NOVASearch (NOVA LINCS, Department of Computer Science Faculty of Science and Technology, University NOVA of Lisbon, Portugal)*;
- NWPU (Northwestern Polytechnical University, China)*;

Participants marked with a star had not participated in the medical task in 2015.

## 3 Results

This section provides the results obtained by the participants in each of the subtasks.

### 3.1 Compound Figure Detection

Table 1 shows the results obtained for the compound figure detection task. Three

**Table 1.** Results of the runs of the compound figure detection task.

| Group | Run | Run Type | Accuracy |
|---|---|---|---|
| DUTIR | CFD_DUTIR_Mixed_AVG | mixed | 92.70 |
| CIS UDEL | CFDRun10 | mixed | 90.74 |
| CIS UDEL | CFDRun09 | mixed | 90.39 |
| CIS UDEL | CFDRun05 | mixed | 90.39 |
| CIS UDEL | CFDRun07 | mixed | 85.47 |
| CIS UDEL | CFDRun8 | mixed | 69.06 |
| CIS UDEL | CFDRun06 | mixed | 52.25 |
| MLKD | CFD2 | textual | 88.13 |
| DUTIR | CFD_DUTIR_Textual_CNN | textual | 87.03 |
| DUTIR | CFD_DUTIR_Textual_RNN | textual | 86.05 |
| CIS UDEL | CFDRun01 | textual | 85.47 |
| DUTIR | CFD_DUTIR_Visual_CNNs | visual | 92.01 |
| CIS UDEL | CFDRun04 | visual | 89.64 |
| CIS UDEL | CFDRun02 | visual | 89.29 |
| CIS UDEL | CFDRun03 | visual | 69.82 |

groups participated in the subfigure detection subtask obtaining an accuracy of up to 92.70% using a multi–modal approach, followed by a visual approach submitted for the same group, DUTIR. DUTIR applied deep convolutional neural networks on vectors trained on the words of all captions using Word2Vec. Five deep convolutional neural networks were also applied on the resized images.

CIS UDEL [11], also achieved its best results using a multi–modal approach. In the textual approach the group extracted a set of delimiters from the captions. In the visual approach, CIS UDEL applied the output of a figure separation approach to classify the figures into compound and non–compound. Finally, the results were fused using several methods, such as a logical union and a decision tree classifier.

MLKD submitted a single run in this subtask achieving the best results using only text information. The textual approach is based on the caption and on the text citing the figure inside the article followed by the use of a random forest classifier.

### 3.2 Figure Separation

Table 2 shows the results for the figure separation subtask. In 2016 only one group participated in the compound figure separation task, CIS UDEL [11], achieving very good results up to an accuracy of 84.43%. Similar results to

**Table 2.** Results of the runs submitted to the figure separation task.

| Run | Group | Run Type | Accuracy |
|---|---|---|---|
| CIS UDEL FS.run9 | | visual | 84.43 |
| CIS UDEL FS.run7 | | visual | 84.08 |
| CIS UDEL FS.run6 | | visual | 84.03 |
| CIS UDEL FS.run8 | | visual | 83.04 |
| CIS UDEL FS.run5 | | visual | 81.23 |
| CIS UDEL FS.run3 | | visual | 75.27 |
| CIS UDEL FS.run4 | | visual | 74.83 |
| CIS UDEL FS.run2 | | visual | 74.30 |
| CIS UDEL FS.run1 | | visual | 73.57 |

2015 were obtained although the difficulty of the subtask was increased in 2016. CIS UDEL applied a connected component analysis to separate the compound figures. A post–processing step was applied to avoid over–fragmentation.

### 3.3 Multi–label Classification

This year two groups submitted runs for the multi–label classification task. The BMET group achieved the best Hamming loss (0.0131) and both groups achieved a F-Measure of 0.32. Table 3 summarises these results.

**Table 3.** Multi–label classification subtasks.

| Run | Group | Hamming Loss | F-Measure |
|---|---|---|---|
| MLC-BMET-multiclass-test-max-all | BMET | 0.0131 | 0.295 |
| MLC-BMET-multiclass-test-prob-max-all | BMET | 0.0135 | 0.320 |
| MLC2 | MLKD | 0.0294 | 0.320 |

The BMET group also submitted a working notes article [12], highlighting the use of Deep Learning and CNNs to classify the images with multiple labels.

### 3.4 Subfigure Classification

This subtask was the most popular task in 2016 with seven groups participating. The results achieved by the participants are shown in Table 4. As in the compound detection task best results were obtained by a multi–modal approach, followed by visual and textual approaches. BCSG [13] achieved the best accuracy of 88.43% by applying multiple visual features and deep convolutional neural networks (CNN). Figure captions and paper full text were also used for the classification. To remove unimportant visual words information gain is used for feature selection. MLKD achieved the best results using a text analysis approach. Similar approaches as in the compound figure detection subtask were applied. Best results on visual approaches were obtained by BCSG followed by IPL [14]. IPL adopted various state–of–the–art visual features, such as, Bag–of–Visual—Words computed with pyramid–histogram–of–visual–word descriptors

**Table 4.** Results of the runs submitted to the subfigure classification task.

| Run | Group | Run Type | Accuracy |
|---|---|---|---|
| BCSG | SC_BCSG_run10_Ensemble_Vote | mixed | 88.43 |
| BCSG | SC_BCSG_run5_Mixed_DeCAF-ResNet-152 | mixed | 88.21 |
| BCSG | SC_BCSG_run4_Hierarchy | mixed | 87.61 |
| BCSG | SC_BCSG_run3_Mixed | mixed | 87.56 |
| BCSG | SC_BCSG_run9_Mixed_NMF | mixed | 86.96 |
| BCSG | SC_BCSG_run6_LateFusion | mixed | 84.44 |
| BCSG | SC_BCSG_run2_Textual | textual | 72.22 |
| MLKD | SC2 | textual | 58.37 |
| BCSG | SC_BCSG_run8_DeCAF_ResNet-152_PseudoInverse | visual | 85.38 |
| BCSG | SC_BCSG_run1_Visual | visual | 84.46 |
| IPL | SC_enriched_GBOC_1x1_256_RGB_Phow_Default_1500_EarlyFusion | visual | 84.01 |
| IPL | SC_enriched_GBOC_1x1_128_HSV_Phow_RGB_1500_EarlyFusion | visual | 83.46 |
| IPL | SC_enriched_GBOC_1x1_128_HSV_Phow_RGB_1500_LateFusion | visual | 82.66 |
| IPL | SC_enriched_GBOC_1x1_128_RGB_Phow_Default_1500_LateFusion | visual | 82.50 |
| IPL | SC_original_GBOC_1x1_256_RGB_w_0.6_Phow_Default_1500_w_0.4_EarlyFusion | visual | 81.73 |
| IPL | SC_original_GBOC_1x1_256_RGB_Phow_Default_1500_EarlyFusion | visual | 81.70 |
| IPL | SC_original_GBOC_1x1_128_RGB_Phow_Default_1500_EarlyFusion | visual | 81.32 |
| BCSG | SC_BCSG_run7_GoogLeNet-PReLU-Xavier | visual | 81.03 |
| IPL | SC_original_GBOC_1x1_256_RGB_Phow_Default_1500_LateFusion | visual | 80.17 |
| IPL | SC_original_GBOC_1x1_128_HSV_Phow_RGB_1500_LateFusion | visual | 80.14 |
| IPL | SC_original_GBOC_1x1_128_RGB_Phow_Default_1500_LateFusion | visual | 79.45 |
| BMET | SC-BMET-subfig-test-prob-sum | visual | 77.55 |
| BMET | SC-BMET-subfig-test-score-sum-merged | visual | 77.53 |
| BMET | SC-BMET-subfig-test-score-sum-cropscale | visual | 77.50 |
| BMET | SC-BMET-subfig-test-majority | visual | 77.26 |
| BMET | SC-BMET-subfig-test-prob-max | visual | 76.38 |
| NWPU | sc.run3 | visual | 74.17 |
| NWPU | sc.run4 | visual | 74.14 |
| NWPU | sc.run5 | visual | 73.97 |
| CIS UDEL | SCRun1 | visual | 72.46 |
| CIS UDEL | SCRun2 | visual | 71.53 |
| NWPU | sc.run2 | visual | 71.41 |
| NWPU | sc.run1 | visual | 71.19 |
| CIS UDEL | SCRun4 | visual | 68.69 |
| CIS UDEL | SCRun3 | visual | 68.17 |
| NOVASearch | SC_NOVASearch_cnn_10_dropout_vgglike.run | visual | 65.31 |
| NOVASearch | SC_NOVASearch_cnn_8_vgglike.run | visual | 65.17 |
| NOVASearch | SC_NOVASearch_cnn_prelu.run | visual | 63.80 |
| NOVASearch | SC_NOVASearch_cnn_10_vgglike.run | visual | 63.29 |
| CIS UDEL | SCRun7 | visual | 53.24 |
| CIS UDEL | SCRun6 | visual | 53.16 |
| CIS UDEL | SCRun5 | visual | 15.62 |

and quad–treebag–of–colors. BMET [12] applied a method similar to the one they used for the multi–label classification task based on CNNs. NWPU also based its method on deep CNNs. A hierarchical approach was used that first classified the figures into diagnostic images and generic biomedical illustrations through a deep CNN. Then, two other deep CNNs were trained to finish the classification of diagnostic images and generic biomedical illustrations,respectively. CIS UDEL [11] also applied a hierarchical classifier using multiple visual descriptors. Neural networks were used as a classifier. Finally, NovaSearch [15] also applied three different CNN models in their approaches.

## 4 Conclusions

In 2016, the ImageCLEF medical task proposed 5 subtasks. One of the subtasks was organized for the first time, the caption prediction subtask. Unfortunately, no participants finally submitted results to the task. This year, more figures were added to the database in the other four subtasks that had already been run in the past. In total, there were eight participants who submitted results, the same number as in 2015 but more runs were submitted in 2016 compared to 2015. The best accuracy obtained was very good in three of the tasks: 92.70% in the compound detection subtask using a multi–modal approach; 84.43% in the figure separation subtask using a visual approach; and 88.43% in the subfigure detection subtask using a multi–modal approach. For the multi–label subtask, the BMET group obtained 0.0135 Hamming loss and a F–Measure of 0.32 using a deep learning approach based on CNNs.

The clear novelty and trend in 2016 is the use of neural network models or deep learning for classification subtasks obtaining very good results in general. CIS UDEL was the only participant of the 2016 figure separation subtask separating the images using a connected component analysis. The main novelty concerning the multi–label task in 2016 was the use of fine–tuned CNNs to perform the multi–label classification.

## Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

## References

1. Clough, P., Müller, H., Sanderson, M.: The CLEF 2004 cross–language image retrieval track. In Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., eds.: Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign. Volume 3491 of Lecture Notes in Computer Science (LNCS)., Bath, UK, Springer (2005) 597–613

2. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., García Seco de Herrera, A., Bromuri, S., Amin, M.A., Kazi Mohammed, M., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., Roldán García, M.d.M.: General overview of ImageCLEF at the CLEF 2015 labs. In: Working Notes of CLEF 2015. Lecture Notes in Computer Science. Springer International Publishing (2015)
3. Müller, H., García Seco de Herrera, A., Kalpathy-Cramer, J., Demner Fushman, D., Antani, S., Eggel, I.: Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: Working Notes of CLEF 2012 (Cross Language Evaluation Forum). (September 2012)
4. García Seco de Herrera, A., Kalpathy-Cramer, J., Demner Fushman, D., Antani, S., Müller, H.: Overview of the ImageCLEF 2013 medical tasks. In: Working Notes of CLEF 2013 (Cross Language Evaluation Forum). (September 2013)
5. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics **39**(0) (2015) 55 – 61
6. García Seco de Herrera, A., Müller, H., Bromuri, S.: Overview of the ImageCLEF 2015 medical classification task. In: Working Notes of CLEF 2015 (Cross Language Evaluation Forum). (September 2015)
7. Bromuri, S., Zufferey, D., Hennebert, J., Schumacher, M.I.: Multi–label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms. Journal of Biomedical Informatics **51** (2014) 165–175
8. Müller, H., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S.: Creating a classification of image types in the medical literature for visual categorization. In: SPIE Medical Imaging. (2012)
9. Bengio, Y.: Deep learning and cultural evolution. In: Genetic and Evolutionary Computation Conference, GECCO, Vancouver, BC, Canada (2014) 1–2
10. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L.: Deep captioning with multimodal recurrent neural networks (m-rnn). CoRR **abs/1412.6632** (2014)
11. Li, P., Sorensen, S., , Kolagunda, A., Jiang, X., Wang, X., Kambhamettu, C., Shatkay, H.: UDEL CIS working notes in ImageCLEF 2016. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, Évora, Portugal, CEUR-WS.org (September 5-8 2016)
12. Kumar, A., Lyndon, D., Kim, J., Dagan, F.: Subfigure and multi–label classification using a fine–tuned convolutional neural network. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, Évora, Portugal, CEUR-WS.org (September 5-8 2016)
13. Koitka, S., Friedrich, C.M.: Traditional feature engineering and deep learning approaches at medical classification task of ImageCLEF 2016. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, Évora, Portugal, CEUR-WS.org (September 5-8 2016)
14. Valavanis, L., Kalamboukis, T.: IPL at CLEF 2016 medical task. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, Évora, Portugal, CEUR-WS.org (September 5-8 2016)
15. Semedo, D., Magalhães, J.a.: NovaSearch at ImageCLEFmed 2016 subfigure classification task. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, Évora, Portugal, CEUR-WS.org (September 5-8 2016)