Branding - Fusion of Meta Data and Musculoskeletal Radiographs for Multi-modal Diagnostic Recognition

Obioma Pelka[1,2], Felix Nensa[2], and Christoph M. Friedrich[1,3]

[1]Department of Computer Science, University of Applied Sciences and Arts Dortmund
{obioma.pelka, christoph.friedrich}@fh-dortmund.de
[2]Department of Diagnostic and Interventional Radiology and Neuroradiology
University Hospital Essen, Germany
felix.nensa@uk-essen.de
[3]Institute for Medical Informatics, Biometry and Epidemiology (IMIBE)
University Hospital Essen, Germany

Abstract

Data fusion techniques provide opportunities for combining information from multiple domains, such as meta and medical report data with radiology images. This helps to obtain knowledge of enriched quality. The objective of this paper is to fuse automatically generated image keywords with radiographs, enabling multi-modal image representations for body part and abnormality recognition. As manual annotation is often impractical, time-consuming and prone to errors, automatic visual recognition and annotation of radiographs is a fundamental step towards computer-aided interpretation. As the number of digital medical images taken daily rapidly increases, there is a need to create systems capable of appropriately detecting and classifying anatomy and abnormality in radiology images. The Long Short-Term Memory (LSTM) based Recurrent Neural Network (RNN) Show-and-Tell model is adopted for keyword generation. The presented work fuses multi-modal information by incorporating automatically generated keywords into radiographs via augmentation. This leads to enriched sufficient features, with which deep learning systems are trained. To demonstrate the proposed approach, evaluation is computed on the Musculoskeletal Radiographs (MURA) using two classification schemes. Prediction accuracy was higher for all tested datasets using the proposed approach with 95.93 % for anatomic regions and 81.5 % for abnormality classification, respectively.

1. Introduction

Due to the fast development in hardware, software, and digital imaging technologies in the medical domain, the amount of information collected per patient scan has rapidly increased [1, 2]. To decrease the burden on radiologists and maintain the maximum interpretation of these radiology reports, the need to implement systems capable of fusing different input domains and thereby obtaining a consolidated data representation, has become more urgent cause. Examples for the medical domain include, incorporating text reports and meta data into radiology images, alongside with combining ECG audio files with medical report findings. There is no restriction to the usage of the fused data, as it can be applied for image classification and annotation [3].

To create effective classification and recognition systems, the selection and fusion of features for a sufficient image representation is essential. As shown in [4, 5, 6, 7, 8], multi-modal representation achieve higher prediction rates in biomedical annotation tasks. The combination of visual and text representation is aimed to sufficiently represent biomedical images. The achieved text features are further adopted for visual recognition tasks, such as body part classification and semantic tagging for substantial structuring of radiographs, as well as abnormality recognition as over 1.7 billion patient worldwide suffer from musculoskeletal problems [11].

However, for real clinical cases and some image classification tasks such as the Musculoskeletal Radiographs (MURA) dataset [9], corresponding text representations are not available. In this paper, the Radiology Objects in COntext (ROCO) dataset [10] is utilized to automatically generate keywords, which are fused with visual features to obtain

multi-modal radiology image representations. Built on the presented work in [8], we extend the fused image representation using a different technique and evaluate the proposed approach for clinical diagnostic recognition.

The presented approach can be adopted to combine different meta data, as well as medical report findings with radiology images to optimize classification prediction accuracy.

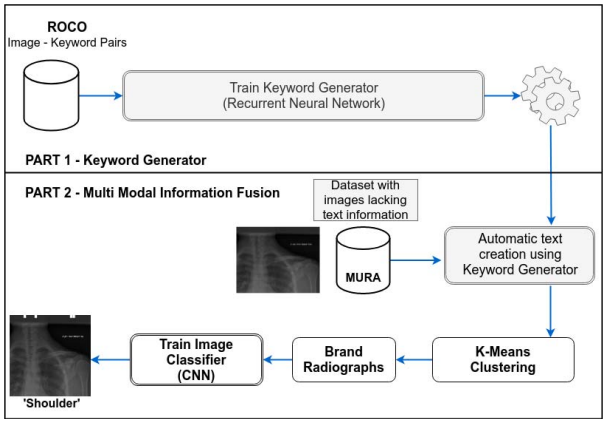


Figure 1. Overview of the proposed approach workflow.

Figure 1 shows the complete workflow for the proposed approach. **PART 1** displays necessary steps for creating a keyword generator and is described in Section 3. This is a distinct and stand-alone process which does not need **PART 2** for application. The keyword generator can be further adopted for several purposes, such as image classification and retrieval.

In **PART 2** detailed in Section 4, the keyword generator is used to create keywords for medical datasets that lack text representations. This second part is dependable on the first part. A medical datasets containing grayscale radiographs and two classification schemes representing different clinical objectives was utilized in **PART 2**, however there is no restriction to x-rays, as the keyword generator was created using radiology images containing a broad variety of modalities.

For biomedical imaging, several early fusion multimodal approaches such as [12, 13, 14] have been presented. These attempts combine image and text representation into one vector, with which the image classifiers are trained on. Adopting this method, the connections in low-level features can be exploited. In [15, 16], late fusion methods were applied, where decision values from several classifiers are fused to make the final classification prediction.

In an attempt combining both early and late fusion advantages, [17] fuse a text document representation with natural images before training the convolutional neural networks for classification. Inspired by this, we propose an approach that brands encoded text features or meta data onto radiographs to obtain an enhanced image representation. The text features used for encoding are word clusters of automatically generated keywords adopting transfer learning. Our contributions in this paper are:

- A novel fusion method by branding radiographs with automatically generated keywords.
- The proposed method is further applied for clinical objective of diagnostic abnormality recognition
- Transfer Learning and LSTM-RNN is utilized for creating textual features for datasets lacking text information.
- Three deep learning based classifiers are trained with the branded radiographs for abnormality recognition and body part classification and labeling.
- The proposed method is evaluated on a radiology dataset with different classification schemes.

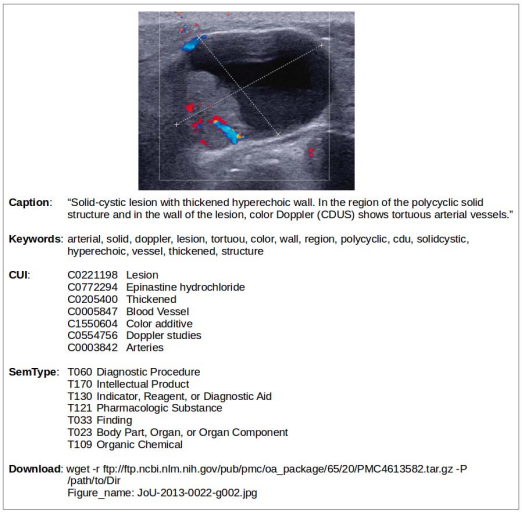
The rest of this paper is structured as follows: Section 2 lists the adopted dataset for keyword generation and the dataset used for evaluating the proposed approach. In Sections 3 and 5, applied deep learning networks, visual representation and machine learning methods for keyword generation and image classification are described. These two sections describe **PART 1** and **PART 2** of Fig. 1. The proposed method applied for incorporating the text features is detailed in Section 4. The achieved results are stated in Section 6. Finally, results are discussed, potential future work and conclusions are drawn in Section 7.

2. Datasets

2.1. Radiology Objects in COntext (ROCO): For Keyword Generation

This dataset is easily accessible, and can be applied for image and information retrieval purposes. Figure 2 shows an example of a radiology image with the corresponding information provided. The Radiology Objects in Context (ROCO) dataset has two classes: Radiology and Out-Of-Class. The first contains 81,825 radiology images, which was used for the presented work, includes several medical imaging modalities such as, Computer Tomography (CT), Ultrasound, X-Ray, Fluoroscopy, Positron Emission Tomography (PET), Mammography, Magnetic Resonance Imaging (MRI), Angiography and PET-CT, and examples of the various modalities can be seen in Fig. 3.

The objective of the ROCO dataset was to provide medical knowledge, originated from peer-reviewed scientific biomedical literature with different textual annotations and a broad scope of medical imaging techniques.



Caption: "Solid-cystic lesion with thickened hyperechoic wall. In the region of the polycyclic solid structure and in the wall of the lesion, color Doppler (CDUS) shows tortuous arterial vessels."

Keywords: arterial, solid, doppler, lesion, tortuou, color, wall, region, polycyclic, cdu, solidcystic, hyperechoic, vessel, thickened, structure

CUI:
C0221198 Lesion
C0772294 Epinastine hydrochloride
C0205400 Thickened
C0005847 Blood Vessel
C1550604 Color additive
C0554756 Doppler studies
C0003842 Arteries

SemType:
T060 Diagnostic Procedure
T170 Intellectual Product
T130 Indicator, Reagent, or Diagnostic Aid
T121 Pharmacologic Substance
T033 Finding
T023 Body Part, Organ, or Organ Component
T109 Organic Chemical

Download: wget -r ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_package/65/20/PMC4613582.tar.gz -P /path/to/Dir
Figure_name: JoU-2013-0022-g002.jpg

Figure 2. Example of a radiology image with corresponding caption, keywords, semantic concepts and types. The ultrasound scan was randomly chosen from the training set of the ROCO dataset [10].

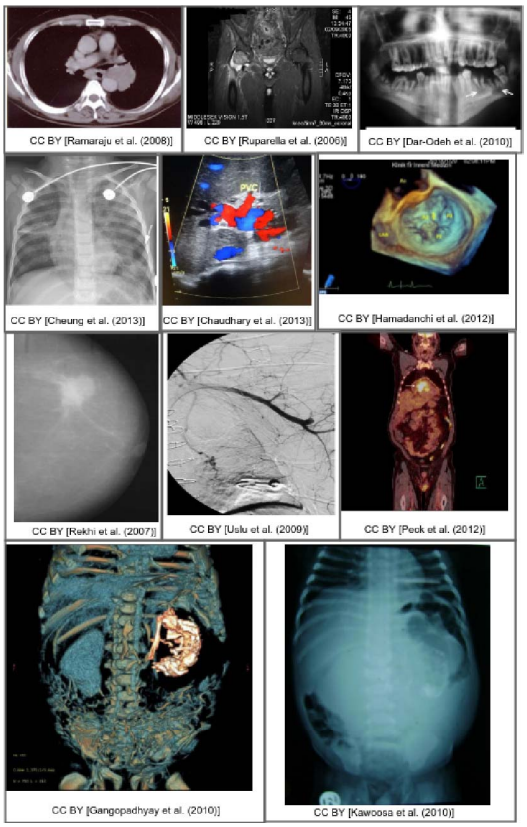


Figure 3. Examples of radiology images contained in the ROCO dataset, illustrating the variety of medical imaging modalities. All images were randomly chosen from the 'Radiology' subset.

From the PubMed [18] Open Access subset, a total number of 6,031,814 image - caption pairs were extracted, which were further for non-compound and radiology images using deep learning systems. Semantic knowledge of object interplay present in the images were extracted in form of UMLS Semantic types and Concept. In addition, the captions were reduced to only nouns and adjectives, which is distributed as Keywords.

2.2. Musculoskeletal Radiographs (MURA) : For Anatomy and Abnormality Classification Evaluation

The MURA dataset of musculoskeletal radiographs presented in [9] consists of 14,863 studies from 12,173 patients. As each study contains one or more views (images), a total number of 40,561 multi-view radiographic images were assembled.

Each of the radiographs belong to one of the seven standard upper extremity radiographic study types: elbow, finger, forearm, hand, humerus, shoulder, and wrist [9]. Figure 4 shows radiographs representing each of the seven anatomy classes. All studies were labeled as normal or abnormal by board-certified radiologists from the Stanford Hospital, between 2011 and 2012, at the interpretation time in the diagnostic radiology environment [9].

Figure 4. Examples of radiographs representing the anatomy classification schemes. The images 'wrist', 'humerus', 'shoulder', 'elbow' and 'finger' belong in addition to the abnormality positive class. Whereas the images 'forearm' and 'hand' belong to the abnormality negative. All images were randomly chosen from the MURA training set [9].

For comparison and research advance purposes, the dataset was split into training (11,184 patients, 13,457 studies, 36,808 images), validation (783 patients, 1,199 studies, 3,197 images), and test (206 patients, 207 studies, 556 images) sets. The explorative analysis regarding class distribution computed on the 9,045 normal and 5,818 abnormal

musculoskeletal radiographic studies is shown in Fig. 5. All images in the training and validation sets are both annotated with corresponding abnormality and body parts labels [9]. For evaluation of the proposed work, the validation set is adopted as the official test set is not public accessible.

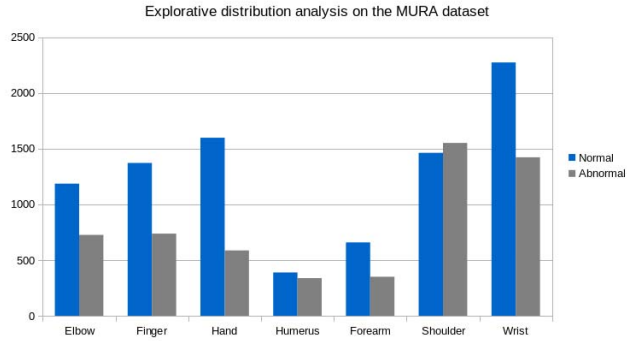


Figure 5. Explorative analysis on the distribution of the Musculoskeletal Radiographs (MURA) dataset. The flow chart shows the total distribution on both training and validation sets.

3. Keyword Generation

As deep learning techniques [19] have improved prediction accuracies in object detection [20], speech recognition [21] and in domain specific applications such as medical imaging [22, 23], a deep learning architecture is used to create the keyword generation model.

Deep Convolutional Neural Networks (dCNN) [24] are applied to encode the medical images to a feature representation which is decoded using a Long Short-Term Memory (LSTM) [25] based Recurrent Neural Network (RNN) [26] to generate appropriate keywords for a given radiograph. This approach, also known as Show-And-Tell model was proposed in [27] and further improved in [28].

To produce rich visual representations of the images, CNN is used as an image encoder by pre-training it for an image classification task. The LSTM-RNN utilized as caption decoder generates the image keywords, using the CNN last hidden layer as input [27].

Figure 6 shows the keyword generation model training setup. In the first training phase, the LSTM is trained using a corpus of paired image and keywords generated from the radiology images in the training set of the Radiology Objects in COntext (ROCO) dataset [10]. No further dataset were used for training. In the second phase, parameters of the image model and LSTM are fine-tuned using the deep learning network Inception-ResNet-V2 [24].

The parameters applied for creating the image keyword generation model are:

- Batch size = [1. Trainingphase = 32; 2. Trainingphase = 32]

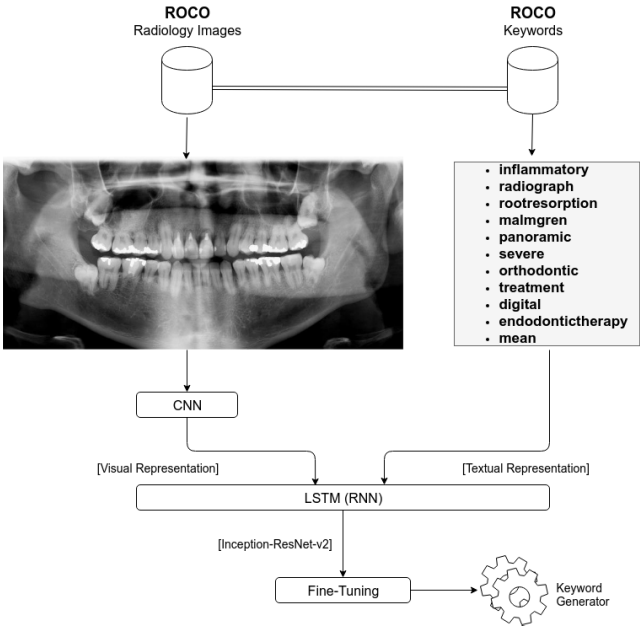


Figure 6. Overview of Long Short-Term Memory based Recurrent Neural Network Model applied for radiology image keyword generation.

- Number of Epochs = [1. Trainingphase = 194; 2. Trainingphase = 583]
- Vocabulary size = 9,750 {Minimum word occurrence ≥ 4}
- Initial learning rate = 2
- Model optimizer = stochastic gradient descent
- Learning rate decay factor = 0.5
- Number of epochs per decay = 8
- Inception learning rate = 0.0005
- Inception model initialization = Inception-ResNet-V2
- LSTM embedding size = 512
- LSTM units number = 512
- LSTM initializer scale = 0.08
- LSTM dropout keep probability = 0.7

For all other parameters not mentioned above, the default values as proposed in [27] and implemented in the Tensorflow-Slim **im2txt**-model [29, 30] were adopted.

Several text preprocessing methods such as reduction of image captions to nouns and adjectives, removal of stopwords [31] and special characters, and word stemming [32] were performed. These text preprocessing steps are further detailed in [33].

4. Radiograph Branding

Utilizing the keyword generation model based on the ROCO dataset and described in subsection 2, keywords were generated for all radiology images in the MURA dataset. Figure 7 show these keywords for two randomly chosen radiographs from the MURA Training Set. No further text preprocessing methods were applied to the generated keywords, as this was done before creating the keyword generation model.

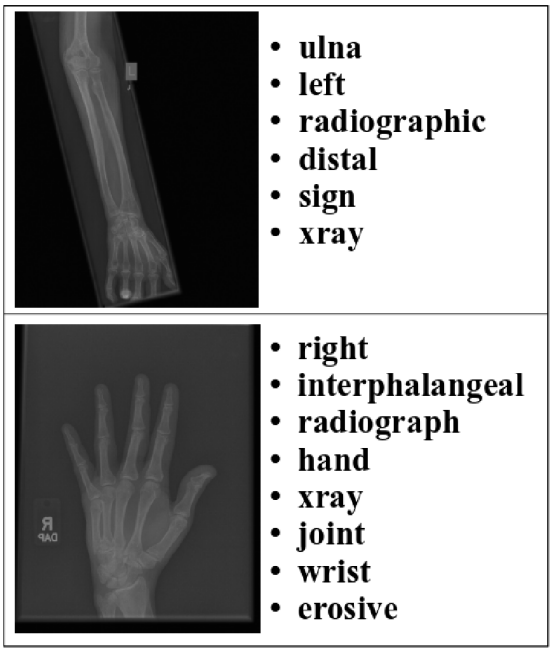


Figure 7. Examples of keywords generated. All images were randomly chosen from the Musculosketal Radiograph MURA Training Set. Keyword generation model was created using the all images from the ROCO radiology class.

Figure 8 shows a subset of the vocabulary obtained with the generated keywords for the MURA dataset. The keywords shown were predicted for images of the abnormality class 'positive'.

Figure 8. Automatically generated keywords for the radiology images in the MURA dataset. All keywords were predicted for images belonging to the abnormality class 'positive'.

To reduce the keywords to semantic concepts, the words in the vocabularies were further grouped to $k = 25$ clusters using the Natural Language Toolkit (NLTK) [31] k-means clustering method [34, 35]. Finally the radiographs (image size [299x299]) are branded by marking the cluster position of the generated keywords on the image, as shown in Fig. 9. The presence of the clusters is incorporated with a [10x10] pixel marker. The complete implementation was done in python.

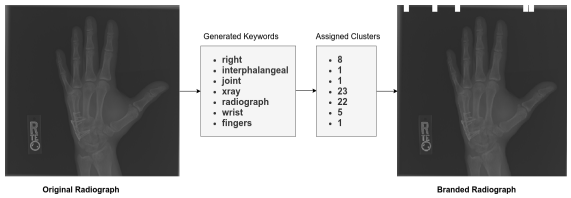


Figure 9. Overview of the complete procedure applied for the radiograph branding. The image was randomly chosen from the MURA training set.

5. Classification

For the dCNNs, TensorFlow-Slim (TF-slim), a lightweight package for defining, training and evaluating models in TensorFlow [29] with pre-trained models, was adopted. To optimize prediction performance, the models were fine-tuned with all trainable weights and best configuration in the second training phase.

Inception-v3 The pre-trained model Inception-v3 [36] which was trained for the ImageNet [37] Large Visual Recognition Challenge 2012 [38], was used to fine-tune the classification model. For the Inception-v3 classification models, the following parameters were applied:

- Optimizer: Root Mean Square Propagation (rmsprop)
- Number of epochs: [1. Trainingphase = 2.5; 2. Trainingphase = 25]
- Number of steps: [1. Trainingphase = 1,000; 2. Trainingphase = 10,000]
- Batch size: 32
- Learning rate: 0.01
- Learning rate decay type: [1. Trainingphase = fixed; 2. Trainingphase = exponential]
- Weight decay: 0.00004
- Model name: Inception-v3

For all other parameters not mentioned above, the default values as proposed in TF-Slim [29] were adopted.

Inception-v4 The pre-trained model Inception-v4 [24] which is a variation of the Inception-v3, having a more uniform simplified architecture and more inception modules, was used to train and fine-tune the second classification model. The following parameters were applied for the Inception-v4 classifier:

- Optimizer: Root Mean Square Propagation (rmsprop)
- Number of epochs: [1. Trainingphase = 2.5; 2. Trainingphase = 25]
- Number of steps: [1. Trainingphase = 1,000; 2. Trainingphase = 10,000]
- Batch size: 32
- Learning rate: 0.01
- Learning rate decay type: [1. Trainingphase = fixed; 2. Trainingphase = exponential]
- Weight decay: 0.00004
- Model name: Inception-v4

For all other parameters not mentioned above, the default values as proposed in TF-Slim [29] were adopted.

Inception-ResNet-v2 The pre-trained model Inception-ResNet-v2 [24] which is a variation of the Inception-v3 using the ideas presented in [39, 40], was used to train and fine-tune the third classification model. For the Inception-ResNet-v2 classification models, the following parameters were applied:

- Optimizer: Root Mean Square Propagation (rmsprop)
- Number of epochs: [1. Trainingphase = 2.5; 2. Trainingphase = 25]
- Number of steps: [1. Trainingphase = 1,000; 2. Trainingphase = 10,000]
- Batch size: 32
- Learning rate: 0.01
- Learning rate decay type: [1. Trainingphase = fixed; 2. Trainingphase = exponential]
- Weight decay: 0.00004
- Model name: Inception-ResNet-v2

For all other parameters not mentioned above, the default values as proposed in TF-Slim [29] were adopted.

Classification Schemes To evaluate the proposed approach, two classification schemes from different datasets were applied. Both classification schemes annotate radiographs according to the body parts examined.

- Musculosketal Radiograph (MURA): Anatomy
  1. Elbow
  2. Finger
  3. Forearm
  4. Hand
  5. Humerus
  6. Shoulder
  7. Wrist
- Musculosketal Radiograph (MURA): Abnormality
  1. Positive
  2. Negative

6. Results

Using the proposed method, increased body parts prediction and abnormality recognition accuracies are obtained on the MURA dataset, which are shown in Table 1 and 2, respectively. Both tables display the performance of the three applied deep learning systems, baseline models, as well as visual, textual and multi-modal results.

The baseline models for the MURA dataset are not comparable with the accuracies in Table 1 and 2, as these were computed using the official test set. For body parts classification, 70.50 % was achieved with a 169-layer DenseNet convolutional neural network [41, 9]. The baseline model for abnormality recognition with 77.80 % was obtained by asking six board-certified radiologists to manually annotate the test images. In addition, Random Forest [42] models were trained with Bag-of-Word (BoW) [43] representations of the generated keywords to show the performance gain of the presented approach.

Table 1. Prediction accuracies obtained using the different visual and text representations and classifier setup, as well as the baseline accuracy presented in [9]. Evaluation was done for body parts classification on Musculosketal Radiographs (MURA) validation set with 3,197 radiographs.

Classifier Setup	Visual	Branded	Textual	Multi
Inception-v3	94.09 %	**95.93 %**	-	-
Inception-v4	92.39 %	94.72 %	-	-
Inception-ResNet-v2	91.84 %	95.00 %	-	-
Random Forest + BoW	-	-	35.11 %	-
Decaf	85.23 %	-	-	-
Decaf + BoW	-	-	-	86.23 %

Table 2. Prediction accuracies obtained using the different visual and text representations and classifier setup, as well as the baseline accuracy presented in [9]. Evaluation was done for abnormality recognition on Musculosketal Radiographs (MURA) validation set with 3,197 radiographs.

Classifier Setup	Visual	Branded	Textual	Multi
Inception-v3	79.85 %	**81.55 %**	-	-
Inception-v4	74.27 %	76.48 %	-	-
Inception-ResNet-v2	78.97 %	79.77 %	-	-
Random Forest + BoW	-	-	54.24 %	-
Decaf	70.98 %	-	-	-
Decaf + BoW	-	-	-	73.20 %

Results obtained with the original radiographs and solely visual features are listed in the first column of the tables. The prediction accuracies obtained with the branded images, denoting the fused text and visual information are shown in the second column and outperform other feature representations. This is observed for both classification schemes. Using the automatically generated keywords without visual representation achieved the poorest prediction rate, which is shown in the third column. The fourth column 'Multi' shows accuracies obtained when combining visual representation from the original image with the automatically generated keywords.

Inception-v3 proved to be the best deep learning system for tackling the two clinical objectives. For body anatomy classification and abnormality recognition, highest accuracy rates were obtained using Inception-v3.

7. Conclusion

This work presents an approach to combine automatically generated keywords with radiographs. Data fusion is achieved by incorporating the textual features by augmentation of the image termed branding. This process enables an enriched multi-modal image representation which is used for body parts classification tasks, as multi-modal image representation has proven to obtain higher prediction results and some image dataset lack text representation. The consolidated multi-modal image representation is further applied for diagnostic recognition of abnormality, as musculoskeletal conditions cause severe and long-time pain.

To create a keyword generation model, image-keywords pairs from the training set of the Radiology Objects in COntext (ROCO) dataset was adopted to train Long Short-Term Memory based Recurrent Neural Network models. Utilizing the keyword generation model, text representations were created for the radiology dataset: Musculosketal Radiograph (MURA), with two classification schemes.

These automatically generated keywords were grouped into k-means clusters and incorporated by augmentation into the radiographs by branding the presence of each cluster in the images.

For both classification schemes, the prediction accuracies obtained with our proposed multi-modal image representation outperformed those achieved just solely visual and textual features, as well other feature fusion methods. The proposed work can be further enhanced by exploiting other word embedding methods, as well as other branding methods, and precedes the way of combining several features of different heterogeneous modalities.

As there are several input sources in the medical domain, the proposed work provides perspective to other data fusion techniques, such as combining meta and medical report findings with other medical imaging modalities.

Acknowledgment

The work of Obioma Pelka was partially funded by a PhD grant from University of Applied Sciences and Arts Dortmund, Germany.

References

[1] M. M. Rahman, P. Bhattacharya, and B. C. Desai, "A Framework for Medical Image Retrieval Using Machine Learning and Statistical Similarity Matching Techniques With Relevance Feedback," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 1, pp. 58–69, 2007. [Online]. Available: https://doi.org/10.1109/TITB.2006.884364 1

[2] H. D. Tagare, C. C. Jaffe, and J. S. Duncan, "Synthesis of research: Medical image databases: A content-based retrieval approach," *Journal of the American Medical Informatics Association JAMIA*, vol. 4, no. 3, pp. 184–198, 1997. [Online]. Available: https://doi.org/10.1136/jamia.1997.0040184 1

[3] M. Ilyas, A. Othmani, and A. Nait-Ali, "Prediction of hearing loss based on auditory perception: A preliminary study," in *First International Workshop on PRedictive Intelligence in MEdicine (PRIME) 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings*, I. Rekik, G. Unal, E. Adeli, and S. H. Park, Eds. Cham: Springer International Publishing, 2018, pp. 34–41. 1

[4] N. Codella, J. Connell, S. Pankanti, M. Merler, and J. R. Smith, "Automated medical image modality recognition by fusion of visual and text information," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*. Springer, 2014, pp. 487–495. 1

[5] L. Valavanis, S. Stathopoulos, and T. Kalamboukis, "IPL at CLEF 2016 Medical Task," in *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, 2016, pp. 413–420. [Online]. Available: http://ceur-ws.org/Vol-1609/16090413.pdf 1

[6] J. Kalpathy-Cramer, A. G. S. de Herrera, D. Demner-Fushman, S. K. Antani, S. Bedrick, and H. Müller,

"Evaluating performance of biomedical image retrieval systems - An overview of the medical image retrieval task at ImageCLEF 2004-2013," *Computerized Medical Imaging and Graphics*, vol. 39, pp. 55–61, 2015. [Online]. Available: https://doi.org/10.1016/j.compmedimag.2014.03.004 1

[7] O. Pelka and C. M. Friedrich, "Modality prediction of biomedical literature images using multimodal feature representation ," *GMS Medizinische Informatik, Biometrie und Epidemiologie*, vol. 12, no. 2, pp. 1345–1359, 2016. [Online]. Available: https://www.egms.de/static/de/journals/mibe/2016-12/mibe000166.shtml 1

[8] O. Pelka, F. Nensa, and C. M. Friedrich, "Variations on branding with text occurrence for optimized body parts classification," in *Proceedings of the 41th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBC 2019, Berlin, Germany, July 23-27, 2019*, 2019. 1, 2

[9] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "MURA dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs," in *Proceedings of the 1st Medical Imaging with Deep Learning, (MIDL) 2018, Amsterdam, Netherlands, July 04-06, 2018.*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Q98pjiG 1, 3, 4, 6, 7

[10] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology objects in context (ROCO): A multimodal image dataset," in *Intravascular Imaging and Computer Assisted Stenting - and - Large-Scale Annotation of Biomedical Data and Expert Label Synthesis - 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings*, 2018, pp. 180–189. [Online]. Available: https://doi.org/10.1007/978-3-030-01364-6_20 1, 3, 4

[11] "The burden of musculoskeletal diseases in the united states," https://www.boneandjointburden.org/print/book/export/html/43. 1

[12] O. Pelka, F. Nensa, and C. M. Friedrich, "Optimizing body region classification with deep convolutional activation features," in *Computer Vision ECCV 2018 Workshops*, ser. Lecture Notes in Computer Science. Springer Nature Switzerland AG, 2019, pp. 1–6. 2

[13] ——, "Adopting semantic information of grayscale radiographs for image classification and retrieval," in *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 2: BIOIMAGING, Funchal, Madeira, Portugal, January 19-21, 2018.*, 2018, pp. 179–187. 2

[14] V. Andrearczyk and H. Müller, "Deep multimodal classification of image types in biomedical journal figures," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings*, 2018, pp. 3–14. [Online]. Available: https://doi.org/10.1007/978-3-319-98932-7_1 2

[15] O. Pelka and C. M. Friedrich, "FHDO Biomedical Computer Science Group at Medical Classification Task of ImageCLEF 2015," in *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*, 2015. [Online]. Available: http://ceur-ws.org/Vol-1391/14-CR.pdf 2

[16] S. Koitka and C. M. Friedrich, "Optimized convolutional neural network ensembles for medical subfigure classification," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction at the 8th International Conference of the CLEF Association, Dublin, Ireland, September 11-14, 2017, Lecture Notes in Computer Science (LNCS) 10456*, G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, and N. Ferro, Eds. Cham: Springer International Publishing, 2017, pp. 57–68. 2

[17] I. Gallo, A. Calefati, S. Nawaz, and M. K. Janjua, "Image and encoded text fusion for multi-modal classification," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA 2018), 10-13 December 2018 in Canberra, Australia*, 2018, pp. 3–14. [Online]. Available: http://artelab.dista.uninsubria.it/res/research/papers/2018/2018-DICTA-Gallo.pdf 2

[18] R. J. Roberts, "PubMed Central: The GenBank of the published literature," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 381–382, Jan. 2001. 3

[19] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. 4

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, USA, July 22-25, 2017*, 2017. 4

[21] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-a. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012. 4

[22] M. S. Abrao, M. O. d. C. Gonçalves, J. A. Dias Jr, S. Podgaec, L. P. Chamie, and R. Blasbalg, "Comparison between clinical examination, transvaginal sonography and magnetic resonance imaging for the diagnosis of deep endometriosis," *Human Reproduction*, vol. 22, no. 12, pp. 3092–3097, 2007. 4

[23] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, and E. I. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 1626–1630. 4

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Resid-

ual Connections on Learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, 2017, pp. 4278–4284. 4, 6

[25] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735 4

[26] Y. Bengio, P. Y. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994. [Online]. Available: https://doi.org/10.1109/72.279181 4

[27] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A neural image caption generator," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 3156–3164. 4

[28] ——, "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 652–663, 2017. 4

[29] Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*. Berkeley, CA, USA: USENIX Association, 2016. 4, 5, 6

[30] C. Shallue. (2018) Im2txt github. [Online]. Available: https://github.com/tensorflow/models/tree/master/research/im2txt 4

[31] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly, 2009. [Online]. Available: http://www.oreilly.de/catalog/9780596516499/index.html 4, 5

[32] M. Porter, "An algorithm for suffix stripping," *Program-electronic Library and Information Systems*, vol. 14, pp. 130–137, 1980. 4

[33] O. Pelka and C. M. Friedrich, "Keyword Generation for Biomedical Image Retrieval with Recurrent Neural Networks," in *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. CEUR-WS Proceedings Notes, Volume 1866*, 2017. 4

[34] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *JSTOR: Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979. 5

[35] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, ser. CVPR '06, 2006, pp. 2169–2178. 5

[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 2818–2826. [Online]. Available: https://doi.org/10.1109/CVPR.2016.308 5

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999134.2999257 5

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015. 5

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition CVPR*. IEEE Computer Society, 2016, pp. 770–778. 6

[40] ——, "Identity mappings in deep residual networks," in *European Conference on Computer Vision ECCV*, ser. Lecture Notes in Computer Science, vol. 9908. Springer, 2016, pp. 630–645. 6

[41] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 2261–2269. 6

[42] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324 6

[43] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, ser. McGraw-Hill computer science series. New York: McGraw-Hill, 1983. 6