NLIP-Essex-ITESM at ImageCLEFcaption 2021 task : Deep Learning-based Information Retrieval and Multi-label Classification towards improving Medical Image Understanding

Janadhip Jacutprakart[1], Francisco Parrilla Andrade[2], Rodolfo Cuan[2], Arely Aceves Compean[2], Giorgos Papanastasiou[1] and Alba G. Seco de Herrera[1]

[1]*University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom*

[2]*Instituto Tecnologico y de Estudios Superiores de Monterrey, Av. Eugenio Garza Sada 2501 Sur, Tecnológico, 64849 Monterrey, N.L., Mexico*

Abstract

This work presents the NLIP-Essex-ITESM team's participation in the concept detection sub-task of the ImageCLEFcaption 2021 task. We developed a method to predict health outcomes from medical images by processing concepts from radiology reports and their associated medical images. Our aim is to improved medical image understanding and provide sophisticated tools to automate the thorough analysis of multi-modal medical images. In this paper, two deep learning- and k-NN-based methods of a) Information Retrieval and b) Multi-label Classification were developed and assessed. In addition, a Densenet-121 and an EfficientNet were used to train and extract imaging features. Our team achieved the second-highest score when the Information Retrieval method was used (F1-score bench-marking was 0.469). Further investigations are underway in the setting of improving health outcome predictions from multi-modal medical images. Code and pre-trained models are available at https://github.com/fjpa121197/ImageCLEF2021.

Keywords

ImageCLEF, image understanding, concept detection, medical image retrieval, Densenet, EfficientNet, k-NN

1. Introduction

This paper presents the contributions of the NLIP-Essex-ITESM team in the ImageCLEFmed caption 2021 task. The team is composed of the Natural Language and Information Processing research group[1] at the School of Computer Science and Electronic Engineering (CSEE) of the University of Essex, and the Instituto Tecnológico y de Estudios Superiores de Monterrey. Since 2003, ImageCLEF [1, 2] held an evaluation campaign as part of the Cross Language Evaluation Forum (CLEF), creating a free online resource on topics and subjects related to cross-language information retrieval. The ImageCLEFmed caption task 2021 edition has two sub-tasks: concept

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

j.jacutprakart@essex.ac.uk (J. Jacutprakart); A00820996@itesm.mx (F. P. Andrade); fitocuan@gmail.com (R. Cuan); arelyac01@gmail.com (A. A. Compean); g.papanastasiou@essex.ac.uk (G. Papanastasiou); alba.garcia@essex.ac.uk (A. G. S. d. Herrera)

[1]https://essexnlip.uk/

detection and caption prediction. The NLIP-Essex-ITESM team participated in 2021 in the concept detection sub-task. The concept detection sub-task focuses on detecting concepts (i.e. UMLS®Concept Unique Identifiers) in a large corpus of radiology images. A detailed description of this year's sub-tasks and data are provided by Pelka et al. [3].

For 2021, we propose two methods: one based on Information Retrieval (IR) and the other based on Multi-label classification (MLC). The IR method used two deep learning models (Densenet-121 [4] and EfficientNet [5]) whilst the MLC used only a Densenet-121 model.

In the ImageCLEFcaption 2020 edition [6], the best results were achieved by the AUEB_NLP team [7]. They examined various Convolutional Neural Network (CNN) models such as ConceptCXN and DenseNet121, combined with two different approaches: feed-forward Neural Network (FFNN) or k-Nearest-Neighbours (k-NN). In this work, we applied the k-NN technique based on the AUEB_NLP team model and referred to our last year submission [8] to evaluate the data and compute the distance. We also implemented the k-NN approach with various metric types to improve the computational time on calculating the distances between a query image and an indexed image to retrieve a similar image in the IR method. Additionally, a new approach using semantic types implemented on MLC.

We developed and implemented several methods in the train and validation dataset. We selected the best approaches based on the top rank information retrieval F1-score evaluation performance. Code and pre-trained models used in this paper are fully publicly available[2].

The paper is structured as follows. Section 2 presents the data collections used in this work. Section 3 details the overall methodology and the two main modelling techniques proposed in this paper (IR and Multi-label classification), including a detailed description of the runs submitted to the ImageCLEFmed caption 2021 task. The results are presented in Section 4 and the conclusions in Section 5.

2. Collection & evaluation

In this work, we used the dataset provided by the ImageCLEFmed caption 2021 task [3]. The dataset consists of:

- *Training set* including 2,756 images-concepts pairs;
- *Validation set* including 500 images-concepts pairs;
- *Test set* including 444 images.

Each image is associated with multiple Unified Medical Language System®(UMLS) Concept Unique Identifiers (CUIs). The UMLS CUIs (associated with the medical images) and 3,256 medical images were included in the training and validation datasets. However, the UMLS CUIs from the test dataset were not distributed where the ImageCLEFcaption task [3] organisers used F1-score to evaluate the submitted runs (see Section 3)

[2]https://github.com/fjpa121197/ImageCLEF2021

3. Methodology

This work proposed two distinct methods: an Information Retrieval (IR)- and a Multi-label classification-based approach. Both methods used two different deep learning models (Densenet-121 and EfficientNet) for image training and feature extraction. In addition, we implemented k-NN to evaluate differences between a query image (whose features were obtained using the same extraction process) and each element in the set with different metrics for the concept selection part. The modelling pipeline uses the preprocessed images in the input and the corresponding concepts in both methods' output.

3.1. Information Retrieval based approach

This approach is based on how a content-based image retrieval system works. Both deep learning models implemented were used for optimally extracting imaging features. The overall approach was separated into three main processes: model training, feature extraction and concept selection.

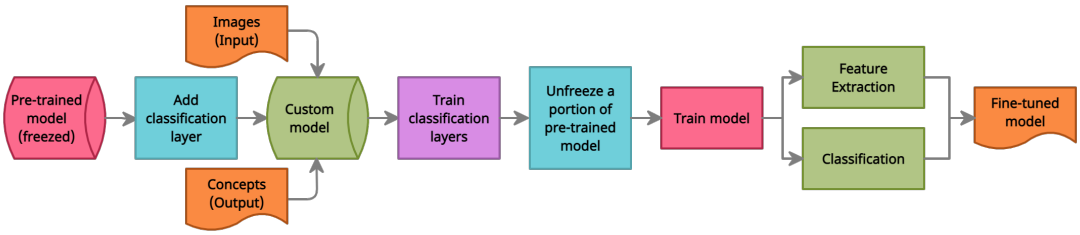

Figure 1: Overview of the model training process.

3.1.1. Model Training.

For this process, both Densenet-121 and EfficientNet models, were used along with pre-trained weights (ImageNet) as base models. For Densenet-121, the input image remained at the maximum input size, 224×224; for EfficientNet, following careful performance evaluations across different input image resolutions, two different input size images were finally selected and examined further: EfficientNet B0 (224×224) and EfficientNet B3 (300×300).

As shown in Figure 1, the process starts by modifying a pre-trained model, Densenet-121 and EfficientNet, and replacing the classification layer with a layer that is allowed to converge to the unique concept output (1585 concepts). In addition, following transfer learning (via fine-tuning) principles, the pre-trained layers of the base model were kept frozen, allowing only the classification model layer to be trained [11]. By default, the activation function used for classification on the last layer uses sigmoid for Densenet-121 [4] and softmax for EfficientNet [5]. The classification layer of the model was first trained for 15 epochs. Once the initial training was over, a portion of the frozen layers was unfrozen, resulting in more trainable layers. The model was then re-trained again for 13 epochs, resulting in a fine-tuned model that was subsequently

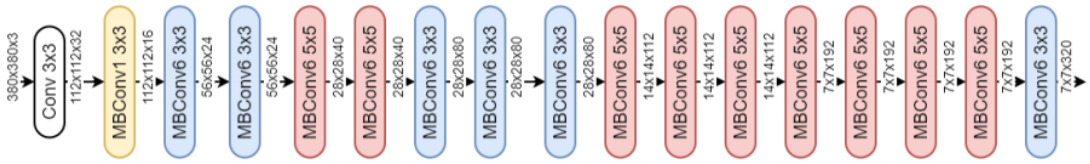

Figure 2: The architecture of EfficientNet B0 [9]

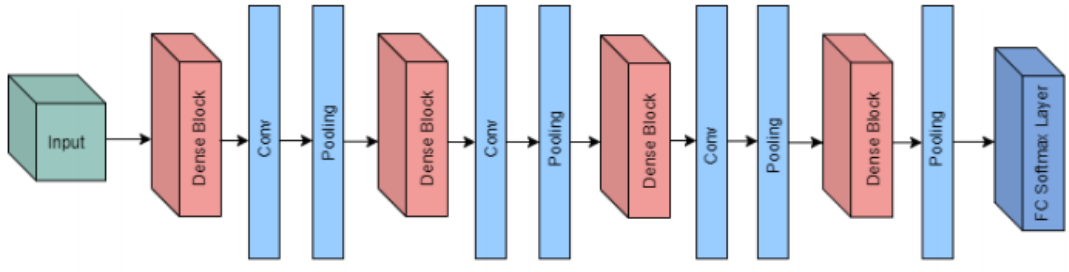

Figure 3: The architecture of DenseNet-121 with 4 dense blocks and 3 transition layers [10]

used in the next step of the process (for feature extraction). The specific parameters used for training are the following:

- *Optimizer*: Adam
- *Learning rate*: 0.001
- *Validation Split* : 0.2
- *Batch size*: 32
- *Loss function*: Binary Cross-Entropy
- *Epochs*: 15 for initial training and 13 for the second training phase

3.1.2. Feature Extraction.

This process consists of using a fine-tuned model to extract features for a set of images. A fine-tuned model (based on Densenet-121 and EfficientNet) was trained as described in the previous process. The same preprocessing steps used in training were followed, and the images were passed to the fine-tuned model. In last year's (ImageCLEF 2020) participation, the output from the batch-normalisation layer was used to get the features for the images. However, following evaluation, we decided that for the Densenet-121, the average pool layer would be used to get the features. On the contrary, EfficientNet used no pooling, which is the default value of the model; as a result, the output of the model will be in the 4D tensor output of the last convolutional layer (sample_size, image width, image height, color_depth). Afterwards, the features were saved to prepare for the next step (the concept selection process, see the following subsection).

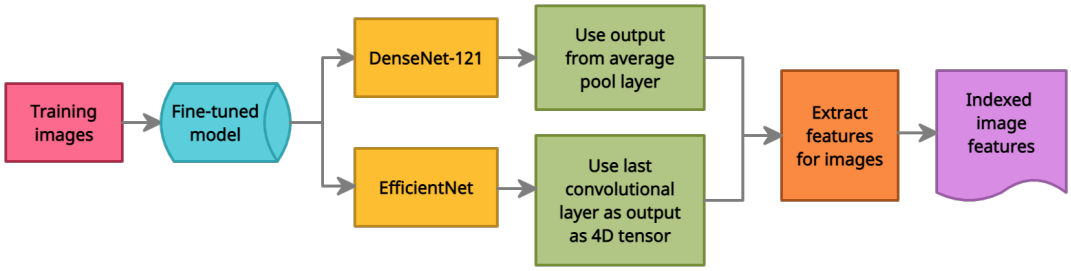

Figure 4: Overview of the feature extraction process.

3.1.3. Concept selection.

For the concept selection process, the set of features obtained from the feature extraction process were indexed by the corresponding imaging data. These indexed features served as the database used to evaluate differences between a query image (whose features were obtained using the same extraction process) and each element in the set. In order to evaluate the differences, distance calculation was implemented using the *k*-NN algorithm. Different experiments were generated using several metrics to find the one that yielded the best results. The distance metrics considered were Canberra, Cosine and Bray–Curtis [12]. We used Bray–Curtis to measured the differences between samples. Based on the F1 score, the Cosine and the Bray–Curtis were the best methods for the Densenet-121 and the EfficientNet, respectively. Although this year's concept selection process had an overall resemblance to our team's method from last year [8], there was also a distance comparison followed by selecting a set of most similar images, where the process was modified to be more efficient. It has generated similar results to the last year in term of computing time using *k*-NN. Along with *k*-NN, through the experiments, a number of different values of k were applied, and we achieved the best outcome with k=1 and using this year's dataset only. As a result, the concepts assigned to the query image correspond to the concepts from the closest indexed image. Since only one image was retrieved, also the ranking process used for last year's concept selection was removed.

3.2. Multi-label Classification

The main characteristic of this approach is that it only uses deep learning to predict the concepts for an image. Since an image can have multiple concepts assigned to it, the problem results in a multi-label classification problem, and consequently, a pre-trained Densenet-121 model has been adapted and used for this task. This approach consists of three main processes: concepts file preprocessing , model training and concepts prediction.

3.2.1. Concepts preprocessing .

In this process, we propose that the concepts which present as output in the competition can be classified by their semantic type (Diagnostic Procedure and Body Part or Organ). We used

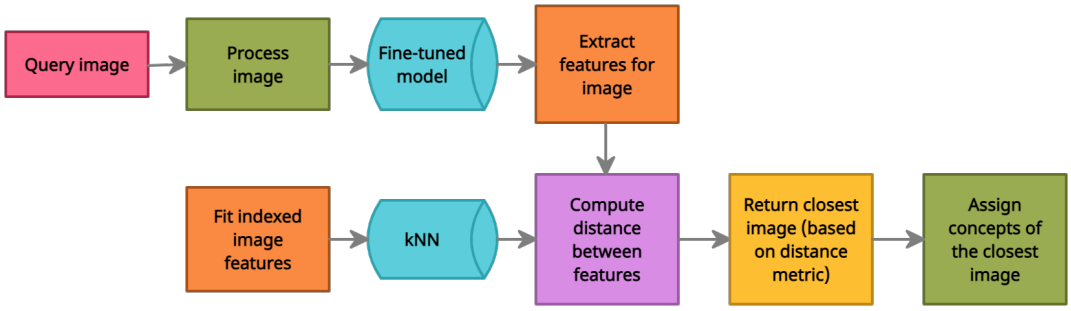

Figure 5: Overview of the concept selection process.

umls-api, which is a Python package in order to retrieve the UMLS REST API. The UMLS REST API and Json output that offer links for important UMLS entities such as CUIs, atoms, and subsets. By using the UMLS REST API, an application programming interface (API) will retrieve a collection of convenient Uniform Resource Identifier (URI) patterns and obtained information related to each concept using its Concept Unique Identifiers (CUI), which consists of the following information: name from the source vocabulary, URLs that refer to the definition and relation(s), date added, semantic type, status, and Unique Identifiers (UI). The python library umls-api[3] was used to access the mentioned API.

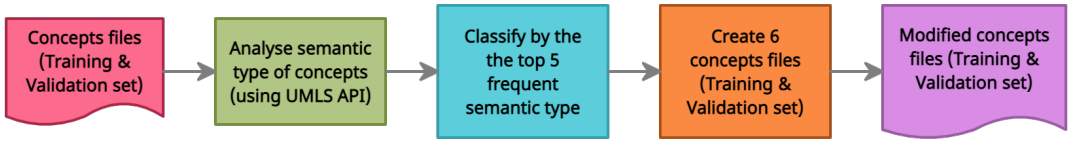

Figure 6: Overview of the preprocessing process.

In order to optimise the number of labels, the semantic type was selected to obtain each concept from the train and validation dataset. As a result, 33 different semantic types were obtained. The following list shows the top 5 most frequent semantic types:

1. Diagnostic Procedure
2. Body Part, Organ or Organ Component
3. Finding
4. Body Location or Region
5. Disease or Syndrome

An individual dataset was made for each type based on the top 5 categories, with an extra

[3]Python library can be found in https://github.com/odwyersoftware/umls-api

dataset for the remaining semantic types. Each dataset consisted of medical images as input and related concepts, based on each image (related to each specific semantic type).

3.2.2. Model training.

As previously mentioned, the preprocessing step creates six new concepts files, each one with the image and its corresponding concepts (of that semantic type). As a result, six models were be trained, where each model corresponds to a specific semantic type. Nonetheless, the model's training process remains the same (except for finding the best threshold).

As illustrated in Figure 7, the process starts by modifying a pre-trained Densenet-121 model (with ImageNet weights) for which the base layers have been frozen. A classification layer is added to the model, which has N number of outputs (this is the unique number of concepts per concept file). The model (only with the classification layer unfrozen) is trained for a certain number of epochs, varying according to each semantic type. Then, a particular portion of the base model is unfrozen, and the model is put into unfrozen again until it exceeds the maximum number of epochs (100) or the callback assigned to it. The specific parameters used for training are the following:

- *Optimiser*: Adam
- *Initial Learning rate used only in classification layer*: 0.0001
- *Learning rate used in second training*: 0.00001
- *Validation Split* : 0.2
- *Batch size*: 32
- *Loss function*: Binary Cross-Entropy
- *Epochs*: This varies depending on the semantic type (see code).

Following two training phases, the output given by the classification layer is a probability (between 0 and 1) that a concept is present given an image; a threshold needs to be found using unseen data. After finding the threshold that gives the best F1-score on unseen data, the model and this threshold will be saved for the concept prediction.

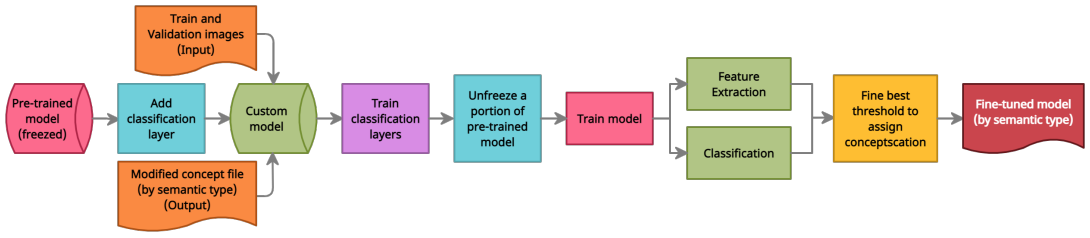

Figure 7: Overview of the training process.

3.2.3. Concepts prediction.

For the prediction of the concepts, the query image passed through the same preprocessing and training method as IR but using only DenseNet-121 model. In this process, we have selected the best threshold found in each model, used only the concepts that passed in that threshold and assigned to the image. In the end, the predictions of each model were merged and used as the final output.

While testing the models and their scores on the validation set, it was noted that using the semantic types result in a positive effect in the overall F1- score; therefore, it was decided only to include these two models when predicting for the test set.

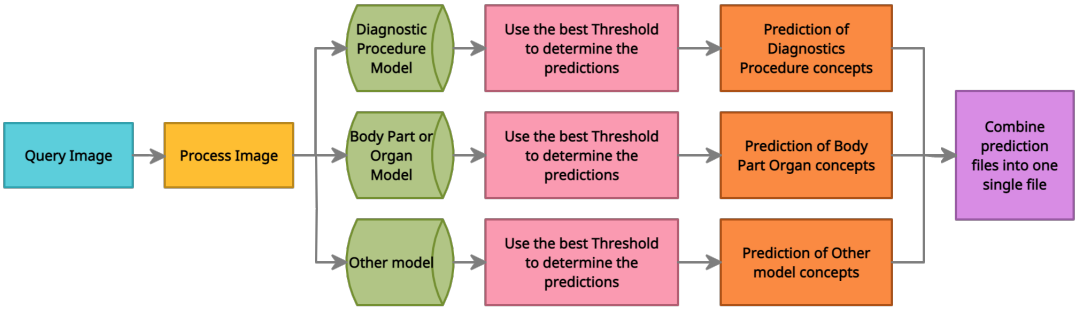

Figure 8: Overview of the concepts prediction process.

3.3. Runs

This section provides a detailed description of the runs submitted to ImageCLEFcaption 2021 task.

- *Run 132945*: For this run, the IR method was implemented, using DenseNet-121 as the feature extraction (average pooling layer) and image inputs are loaded as 224×224. The length of the resulting feature vector for each image is 1024. Then, (k-NN) (k = 1 and metric = cosine) is used to retrieve the closest image and the concepts of the closest image are assigned to the query image.
- *Run 136379*: For this run, the IR method is implemented, using EfficientNet B3 as feature extraction (no pooling layers) and image inputs are loaded as 300×300. The (k-NN) (k = 1 and metric = canberra) is used to retrieve the closest image and the concepts of the closest image are assigned to the query image.
- *Run 136400*: In this run, the IR method is implemented, using EfficientNet B0 as feature extraction (no pooling layers) and image inputs are loaded as 224×224. The (k-NN) (k = 1 and metric = canberra) is used to retrieve the closest image and the concepts of the closest image are assigned to the query image.
- *Run 136404*: For this run, the IR method is implemented, using EfficientNet B0 as feature extraction (no pooling layers) and image inputs are loaded as 224×224. The (k-NN) (k =

1 and metric = cosine) is used to retrieve the closest image and the concepts of the closest image are assigned to the query image.

- *Run 136429*: For this run, the IR method is implemented, using EfficientNet B0 as feature extraction (no pooling layers) and image inputs are loaded as 224×224. The (k-NN) (k = 1 and metric = braycurtis) is used to retrieve the closest image and the concepts of the closest image are assigned to the query image.
- *Run 133912*: For this run, the MLC method is implemented, and based on F1-score obtained from the validation set, only the models for the diagnostic procedure and body part or organ were included in the finals predictions. The threshold used for assigning a concept for the diagnostic procedure model was 0.4 and for the body part or organ was 0.1. The other semantic types were not included because when testing on the validation set, adding the additional semantic types affected negatively the overall F1-score.

4. Results

Table 1 summarises the techniques used by each run. Table 1 presents the official results our

Table 1
Description and performance of the runs submitted to ImageCLEF 2021 Concept Detection Task and their ranks compared with all the 29 runs submitted by the 5 participating teams.

Run ID	Size Input Image	Method	DL Model	Similarity measure	F1-Score	Ranking
132945	224×224	IR	Densenet-121	Cosine	**0.469**	6
133912	224×224	Multi-label classification	Densenet-121	N/A	**0.412**	15
136379	300×300	IR	EfficienNetB3	Canberra	**0.355**	21
136400	224×224	IR	EfficienNetB0	Canberra	**0.423**	13
136404	224×224	IR	EfficienNetB0	Cosine	**0.440**	12
136429	224×224	IR	EfficienNetB0	BrayCurtis	**0.451**	11
Best ImageCLEF2021		-	-	-	**0.505**	1

team achieved in the ImageCLEF 2021 Concept Detection Task and ranking along other 29 runs submitted by the five different teams. Our team received the second place in overall score with the best results on *132945* with F1-score of 0.469, close to a couple of submission from the first place team, AUEBs_NLP_Group who achieved F1-score of 0.505, 0.495, 0.493, 0.490, respectively.

Notably, our best submission used DenseNet-121 as the feature extraction (average pooling layer)and image inputs are loaded as 224×224 for fine-tuning and image extraction with IR method using k-NN method with Cosine distance to retrieve the closest image and the concepts of the closest image are assigned to the query image. Based on the final results that our team has achieved, it is clear that using the further IR method with different distance metrics also improve the score differently between DenseNet-121 and EfficientNet. However, with another method using Multi-label classification, we use the same DenseNet-121 with the IR method but with a different process in the latter part. It might be due to the size of the dataset, in which a bigger dataset is required in order for the Multi-label classification method to be efficient. Besides the two different deep learning models we used this year, there is a slight difference in both results from using DenseNet-121 and EfficientNet. Similarly, using three different similarity

metrics (Canberra, Cosine and BrayCurtis) also resulting in a slightly best score using Cosine on DenseNet-121 and BrayCurtis in the EfficientNet model.

5. Conclusions

This paper describes our contributions in the ImageCLEFcaption 2021 task. Two different methods were developed and used in this paper. An information retrieval method using two deep learning models, a DenseNet-121 and an EfficientNet ,to train and extract features from the data collections. At the same time, multi-label classification was implemented using a DenseNet-121 only. Considering the baseline model from last year [8], we have differentiated and optimised our modelling pipeline to further generalise our approach and improve outcomes. Our DenseNet-121 model showed the highest performance when incorporated in IR method. Followig this method, we achieved the second-best performance (F1-score of 0.469. Unlike the previous year, no additional modality information was provided, which added additional complexixity in our processing pipeline. Further investigations on developing and customising deep learning model arcitectures and fine-tuning are already underway, so that we will further improve model performannce.

References

[1] J. Kalpathy-Cramer, A. García Seco de Herrera, D. Demner-Fushman, S. Antani, S. Bedrick, 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 39 (2015) 55–61.

[2] B. Ionescu, H. Müller, R. Peteri, A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, S. Kozlovski, V. Liauchuk, Y. Dicente, V. Kovalev, O. Pelka, A. G. S. de Herrera, J. Jacutprakart, C. M. Friedrich, R. Berari, A. Tauteanu, D. Fichou, P. Brie, M. Dogariu, L. D. Ştefan, M. G. Constantin, J. Chamberlain, A. Campello, A. Clark, T. A. Oliver, H. Moustahfid, A. Popescu, J. Deshayes-Chossart, Overview of the ImageCLEF 2021: Multimedia retrieval in medical, nature, internet and social media applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021), LNCS Lecture Notes in Computer Science, Springer, Bucharest, Romania, 2021.

[3] O. Pelka, A. Ben Abacha, A. García Seco de Herrera, J. Jacutprakart, C. M. Friedrich, H. Müller, Overview of the ImageCLEFmed 2021 concept & caption prediction task, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.

[4] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[5] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[6] B. Ionescu, H. Müller, R. Péteri, A. B. Abacha, V. Datla, S. A. Hasan, D. Demner-Fushman,

S. Kozlovski, V. Liauchuk, Y. D. Cid, V. Kovalev, O. Pelka, C. M. Friedrich, A. G. S. de Herrera, V.-T. Ninh, T.-K. Le, L. Zhou, L. Piras, M. Riegler, P. l Halvorsen, M.-T. Tran, M. Lux, C. Gurrin, D.-T. Dang-Nguyen, J. Chamberlain, A. Clark, A. Campello, D. Fichou, R. Berari, P. Brie, M. Dogariu, L. D. Ştefan, M. G. Constantin, Overview of the ImageCLEF 2020: Multimedia retrieval in medical, lifelogging, nature, and internet applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 12260 of *Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020)*, LNCS Lecture Notes in Computer Science, Springer, Thessaloniki, Greece, 2020.

[7] B. Karatzas, J. Pavlopoulos, V. Kougia, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption 2020, in: CLEF 2020 Working Notes, Thessaloniki, Greece, September 22-25, 2020, 2020.

[8] A. G. S. de Herrera, F. P. Andrade, L. Bentley, A. A. Compean, Essex at Image-CLEFcaption 2020 task, in: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS. org, Thessaloniki, Greece (September 22-25 2020), 2020.

[9] T. A. Putra, S. I. Rufaida, J.-S. Leu, Enhanced Skin Condition Prediction Through Machine Learning Using Dynamic Training and Testing Augmentation, IEEE Access 8 (2020) 40536–40546.

[10] L. Sarker, M. M. Islam, T. Hannan, Z. Ahmed, Covid-densenet: A deep learning architecture to detect covid-19 from chest radiology images (2020).

[11] P. Dutta, P. Upadhyay, M. De, R. Khalkar, Medical image analysis using deep convolutional neural networks: CNN architectures and transfer learning, in: 2020 International Conference on Inventive Computation Technologies (ICICT), IEEE, 2020, pp. 175–180.

[12] N. Thakur, D. Mehrotra, A. Bansal, M. Bala, Analysis and Implementation of the Bray–Curtis Distance-Based Similarity Measure for Retrieving Information from the Medical Repository, in: International Conference on Innovative Computing and Communications, Springer, 2019, pp. 117–125.