Was that so hard? Estimating human classification difficulty

Morten Rieger Hannemose[*,**1], Josefine Vilsbøll Sundgaard[*1], Niels Kvorning Ternov[2], Rasmus R. Paulsen[1], and Anders Nymark Christensen[1]

[1] Department of Applied Mathematics and Computer Science,
Technical University of Denmark, Kgs. Lyngby, Denmark
[2] Department of Plastic Surgery, Copenhagen University,
Herlev and Gentofte Hospital, Copenhagen, Denmark

Abstract. When doctors are trained to diagnose a specific disease, they learn faster when presented with cases in order of increasing difficulty. This creates the need for automatically estimating how difficult it is for doctors to classify a given case. In this paper, we introduce methods for estimating how hard it is for a doctor to diagnose a case represented by a medical image, both when ground truth difficulties are available for training, and when they are not. Our methods are based on embeddings obtained with deep metric learning. Additionally, we introduce a practical method for obtaining ground truth human difficulty for each image case in a dataset using self-assessed certainty. We apply our methods to two different medical datasets, achieving high Kendall rank correlation coefficients on both, showing that we outperform existing methods by a large margin on our problem and data.

Keywords: Difficulty estimation · Deep metric learning · Human classification.

1 Introduction

When doctors diagnose patients, not all cases have the same diagnostic difficulty. A case can be very easy if there are clear diagnostic signs. However, if the typical signs are missing or give conflicting information, a doctor will be more likely to assign an incorrect diagnosis. When doctors are trained to diagnose certain diseases, they learn faster when starting with easy cases and then gradually progressing to harder cases [22]. Knowing how hard each case is to classify is thus useful in an educational context. This concept is well-known in pedagogy [9] and applies to many other areas such as language training, mathematics, etc.

In this paper, we present a novel approach for estimating human classification difficulty using deep metric learning. In deep metric learning, high-dimensional

---

[*] These authors contributed equally
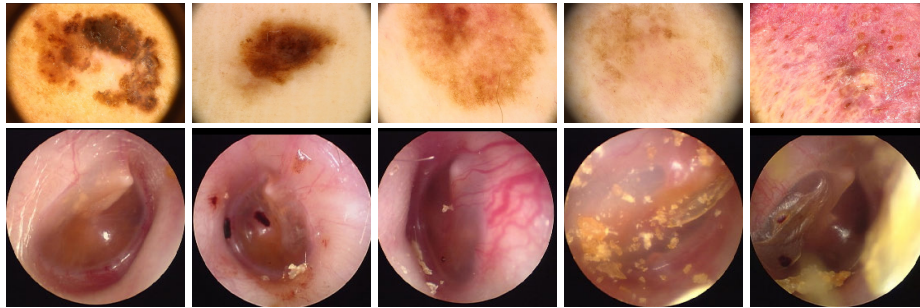[**] Corresponding author

Fig. 1. Image examples from the skin lesion (top row) and eardrum (bottom row) datasets. The difficulty increases from left to right from 0 to 1 in steps of 0.25 for each image. For the skin lesion dataset, only images from the melanoma class are shown, while the eardrum images are from all three diagnostic classes, see Section 3.

data (in our case, images) are mapped to a lower-dimensional embedding that captures similarities between the training examples: Similar images cluster together, and dissimilar images are pushed apart. In our paper, we define metrics in the embedding space that capture human classification difficulty. We evaluate our methods on two different medical datasets, one containing images of skin lesions and the other of eardrums, see Fig. 1.

The term *difficulty* is used in various ways in image analysis. Difficulty can be defined as how hard it is for machine learning to reach high accuracy on a given dataset [25], how challenging it is to automatically segment an image [16], visual complexity and clutter in the image [19], the time needed for a human to segment an image [33], or the human response time for a visual search task [31]. The latter definition was employed by Ionescu *et al.* [31], who proposed a method based on a pretrained neural network for feature extraction, followed by support vector regression to estimate the difficulty score. They presented a dataset with difficulty scores on the PASCAL VOC2012 dataset evaluated by 736 raters.They achieved a high Kendall's τ coefficient of 0.472. In contrast to our approach, they do not use any knowledge about the ground truth class of the image but instead estimate the difficulty directly from image features. By using both the ground truth class and an embedding space our approach becomes more interpretable [24]. We define human difficulty as the fraction of incorrect classifications from people familiar with the classification task. This definition was also employed by André *et al.* [1]. For one of our datasets, we use a self-evaluated certainty of all raters to obtain a less noisy estimated ground truth difficulty with few annotators.

We compare our work to methods in active learning and curriculum learning. Active learning accelerates labeling efficiency by selecting the most useful samples from an unlabeled dataset for labeling, thus reducing the labeling cost [21]. The intuition behind the most commonly used approach, the uncertainty-based approach, is that with a lower certainty on a specific example, a higher amount of informativeness will be added to the classifier when utilizing the example for training [35]. Curriculum learning is inspired by the learning process of humans,

where examples are presented with increasing order of difficulty. This concept is transferred to neural networks to increase training speed and performance by introducing easy examples at the beginning of training, and to gradually increase the difficulty of the training examples [2].

In this paper, we present a new procedure for obtaining ground truth human difficulty from several annotators by including a self-evaluated certainty. We also propose a new method for estimation of human difficulty based on embeddings of images learned using deep metric learning, which outperforms existing methods by a large margin. We propose methods that both utilize ground truth difficulties for training and methods that do not. Finally, we are the first to include the ground truth class label for human difficulty estimation, which increases the performance of our methods even further.

2   Estimating image difficulty

Our difficulty estimation models are all based on the embedding space learned using a deep neural network, trained with metric learning. By training a model this way, instead of as a classification network, we learn the similarities in the training dataset. The output from the network is an embedding vector, mapping each individual image to the embedding space. The idea is that easy cases will be placed far from decision boundaries in the embedding space, while difficult cases will be further away from the class cluster center, and possibly closer to other cluster classes. We separate our proposed methods into two categories depending on whether or not they utilize ground truth difficulties during training. An overview of these methods is in Table 1.

Methods without ground truth difficulties are all based on embeddings of samples, extracted using a trained neural network. As our neural networks are trained using cosine similarity, our methods for estimating difficulties are thus also based on cosine similarities. As difficulties should be high for points far from their cluster, we refer to inverse similarity which is one minus the similarity. The methods still apply to neural networks trained using Euclidean distances, and in that case, one would use the Euclidean distance in the embedding space instead.

Inverse similarity is a naïve approach to estimating the difficulty, found by computing the similarity between the sample and the cluster center of its ground truth class. This is intuitive, as samples less similar to the cluster center are typically more similar to other class clusters, and thus harder to classify. To find the difficulty, and not the easiness, we report the inverse of the similarity.

Inverse softmax of similarity is an improvement of inverse similarity. Samples can have low similarity to their cluster center without being close to other classes. To handle this, we compute the similarity between the sample and all cluster centers and normalize these with softmax. The difficulty is the inverse of the softmax output corresponding to the ground truth class. This method is related to decision margin sampling in active learning [32], except we can go on both sides of the decision boundary since the ground truth *class label* is known.

Sample classification power is an alternative way of obtaining an estimate of image difficulty. Here, we evaluate how many of the neighboring points in the embedding space belong to the ground truth class of a certain sample. To do that for a single sample s from class c, we imagine classifying the closest k samples as c, and classifying the rest as not c. By varying k from one to the number of samples, we can draw a receiver operating characteristic (ROC) curve. We then use the area under the curve (AUC) of this ROC curve as our estimate of the difficulty of s. To handle class imbalance, we use the weighted ROC curve, with the weights being the inverses of the class frequencies.

Normalization is carried out on the estimated difficulties, by introducing the assumption that each class has the same average difficulty. To enforce this assumption, we propose normalizing the difficulty on a per-class basis by dividing it by the average estimated difficulty of that class. We refer to this as "norm".

Methods with ground truth difficulties are methods, where the ground truth difficulties of a training set are employed. These can also be called supervised methods. We set up a regression problem to predict the difficulty scores directly from the image embeddings. We employ the tree-based ensemble model extra trees [10] for the regression problem. In addition to only predicting from the embeddings, we also fit a model using the ground truth label as additional input. The ground truth label will allow the model to learn that samples placed close to incorrect class clusters should have a higher difficulty, than samples within their correct class cluster. Several other models were tested for this task, including support vector regression, but extra trees showed superior performance.

3   Datasets

To validate our method, we performed experiments on two medical image datasets, examples of which are shown in Fig. 1. We have obtained estimates of the human difficulty for a number of images from both datasets, which we use as our test-sets for evaluating our proposed approaches.

The skin lesion dataset consists of dermoscopic images of skin lesions divided into eight diagnoses, which include benign (nevus [NV], keratoses [BKL], vascular lesions [VASC], dermatofibromas [DF]), pre-malignant (actinic keratoses), and malignant (melanoma [MEL], squamous cell carcinoma [SCC], basal cell carcinoma [BCC]). The diagnoses were determined by histopathology or as the consensus between two to three domain experts. We have a dataset of 52 292 images from the 2019 ISIC Challenge training set [30,5,6] [3] and our own dataset (Permission to access and handle the patients' data was granted by the Danish Patient Safety Authority (Jr.# 3-3013-2553/1) and the Data Protection Agency of Southern Denmark (Jr.# 18/53664)).

Skin lesion difficulties are obtained for 1723 images from our own dataset, based on diagnoses from 81 medical students with an interest in dermatology (Ethical waiver: Jr.#: H-20066667, data handling agreement case #: P-2019-556). On average, each student diagnosed 609 randomly sampled images. It was

---

[3] License: CC-BY-NC

ensured that at least eight students diagnosed each case. The images were diagnosed into seven different categories, as we expected actinic keratoses would be too difficult for the medical students. We estimate the difficulty of a case as the fraction of students answering incorrectly.

The eardrum dataset contains 1409 images collected during the standard clinical routine at an Ear-Nose-and-Throat (ENT) clinic. The data was collected under the ethical approval from the Non-Profit Organization MINS Institutional Review Board (ref.# 190221). The images show the patients' eardrum captured using an endoscope and are diagnosed into three different diagnoses: acute otitis media, otitis media with effusion, and no effusion by an experienced ENT specialist. The dataset is split into a training and test set of 1209 and 204 images.

Eardrum difficulties were estimated by getting the test set of 204 equally class sampled eardrum images analyzed and diagnosed by four additional experienced ENTs. The ENTs diagnosed each case as one of the three diagnoses or "unknown", counting as an incorrect diagnosis. Furthermore, each ENT rated their certainty of each diagnosis on the scale: very low, low, medium, moderate, or high, which is converted to a scale from 0 to 1. More details on this dataset are in Sundgaard *et al.* [28]. For a case, $\mu_{correct}$ is the fraction of correct ENT answers and $\mu_{certainty}$ is the average self-evaluated certainty. The difficulty of each case is then

$$1 - \mu_{correct} \cdot \mu_{certainty}. \tag{1}$$

We evaluate the eardrum difficulties with "leave-one-annotator-out", by comparing the responses from one annotator with the difficulties estimated from the remaining annotators using Kendall's τ. This resulted in an average Kendall's τ of 0.548 based the difficulty using only on the fraction of correct ENT answers. The score increased to 0.570, corresponding to 78.5% correctly ranked cases, when including the self-evaluated certainty in the definition of difficulty, showing that this improves the estimated difficulties.

4 Experiments

The embeddings of the images in our proposed methods are computed using neural networks trained with a metric loss function. All experiments are conducted in PyTorch (v. 1.10) using the PyTorch metric learning library [18]. The neural networks are trained using the multi-similarity loss function [34] ($\alpha = 2$, $\beta = 50$, base = 1) and a multi similarity miner ($\epsilon = 0.1$) using cosine similarity to optimize the selection of training pairs. Our models are pretrained on the ImageNet database [8]. The fully connected layer before the final softmax of the model is replaced by a fully connected layer without an activation function, which returns the embedding space. The output embeddings are L2 normalized.

The skin lesion network is based on a ResNet-50 model [12], with a 64-dimensional embedding space. The model is trained for 350 epochs with a learning rate of 10^{-5}. The input images (256×256) are color normalized using the Minkowski norm ($p = 6$). Data augmentation consists of flips, rotations, scaling, and color jitter. We do inference with the same augmentations, and compute each

prediction as the average of 64 random augmentations. The eardrum network is based on the Inception V3 network [29], with a 32-dimensional embedding space. The Inception V3 network has been used by several others for similar images [4,27]. The parameters of the first half of the network (until first grid size reduction) were frozen to avoid over-fitting. The initial learning rate (10^{-3}) is decreased by a factor of 0.1 every 50th epoch. Training is continued until the training loss has not decreased for 20 epochs, resulting in 111 training epochs. Data augmentation consists of horizontal flips, rotations, color jitter, and random erasing. Images are resized to 299×299.

We use Kendall's τ [13] to evaluate how well our methods can predict the ground truth difficulties. This is a non-parametric measurement from -1 to 1 of the correlation between two ranked variables. As it only compares how the images are ranked, it is not important to achieve the exact same difficulty as the ground truth estimate, as long as the ordering of samples is correct.

We use Extra trees [10] for supervised difficulty estimation, with five-fold cross-validation. This allows us to obtain predictions for all samples in the test set, and thus compute a single Kendall's τ for the entire test set. All our experiments with extra trees use 500 trees, with 10 as the minimum number of samples required to split an internal node.[4]

Comparisons are made between our methods and methods from both active and curriculum learning using a standard trained classification network, and with the approach proposed by Ionescu *et al.* [31]. The classification networks employ the same architecture as our embedding networks, but the dimension of the output is the number of classes in each dataset. The networks are trained with cross-entropy loss weighted by the inverse frequency of each class, but otherwise using the same setup as described for the embedding networks.

Visual search difficulty proposed by Ionescu *et al.* [31] is used for comparison. We replicate their method by passing each image (299×299) through VGG-16 [26] once and using the penultimate features to fit a ν-support vector regression.[4]

We compare to the following approaches from active learning, all based on the softmax output of a classification network: classification uncertainty, which is one minus the maximum value of the softmax [14]; entropy of the softmax probabilities [7]; and classification margin found by computing the difference between the second-highest and highest probabilities of the softmax [15].

We also compare to three approaches from curriculum learning: standard deviation of the images [23]; transfer scores obtained by running all images through a pretrained Inception V3 network using the penultimate features to train a support vector classifier to obtain the confidence of the model [11]; and one minus the softmax output of the ground truth class from our classification network [11]. These active and curriculum learning methods are only used to estimate the difficulty for each image, and have not been utilized to train the neural networks.

---

[4] Unspecified parameters are the defaults in Scikit-Learn v. 0.24.2 [20].

Table 1. Kendall's τ for all methods on both datasets. Methods with (L) use the ground truth class label for prediction, and with (D) use a training set of ground truth difficulties. Bold indicates the significantly best performance, and bold with a star indicates the methods without D performing significantly better than the rest. We used nonparametric bootstrap based testing [3] with 50 000 replicates and $\alpha = 5\%$.

Method	Uses		Skin lesion	Eardrum
Visual search difficulty [31]		D	0.142	0.117
Curriculum learning				
Std. of image [23]			−0.070	0.011
Transfer scoring [11]	L		0.115	0.213
Self-taught scoring [11]	L		0.176	0.261
Active learning				
Classification uncertainty [14]			0.094	0.217
Entropy of probabilities [7]			0.118	0.216
Classification margin [15]			0.068	0.215
Ours				
Inverse similarity	L		0.137	−0.140
Inverse softmax of similarity	L		**0.239***	0.354
Inverse softmax of similarity norm.	L		**0.239***	0.380
Sample classification power	L		0.201	0.143
Sample classification power norm.	L		**0.247***	**0.440***
Extra trees: embeddings		D	0.322	0.465
Extra trees: embeddings + label	L	D	**0.398**	**0.517**

5   Results

The Kendall's τ for all experiments is reported in Table 1. The table also gives an overview of whether the ground truth label is used for prediction, and whether a training set of ground truth difficulties has been used. The embeddings for the two datasets are shown in the top of Figure 2. For the eardrum data, we see how most easy examples are located within the class clusters, while the difficult examples are the ones located in another class cluster, or at the edge of the clusters. The same tendencies are visible in some classes of the skin lesion embeddings. The bottom of Figure 2 shows scatter plots of the ground truth difficulties versus predicted difficulties for both datasets.

6   Discussion and conclusion

We have shown that image embeddings from neural networks trained with metric learning can be used to estimate diagnostic difficulty. Our methods for difficulty estimation outperform all existing methods in both active and curriculum learning. Ionescu *et al.* [31] report a Kendall's τ of 0.472, while their method achieves 0.142 and 0.117 on our datasets. Their method is, however, developed for a different problem, thus this drop in performance is not surprising. Our methods
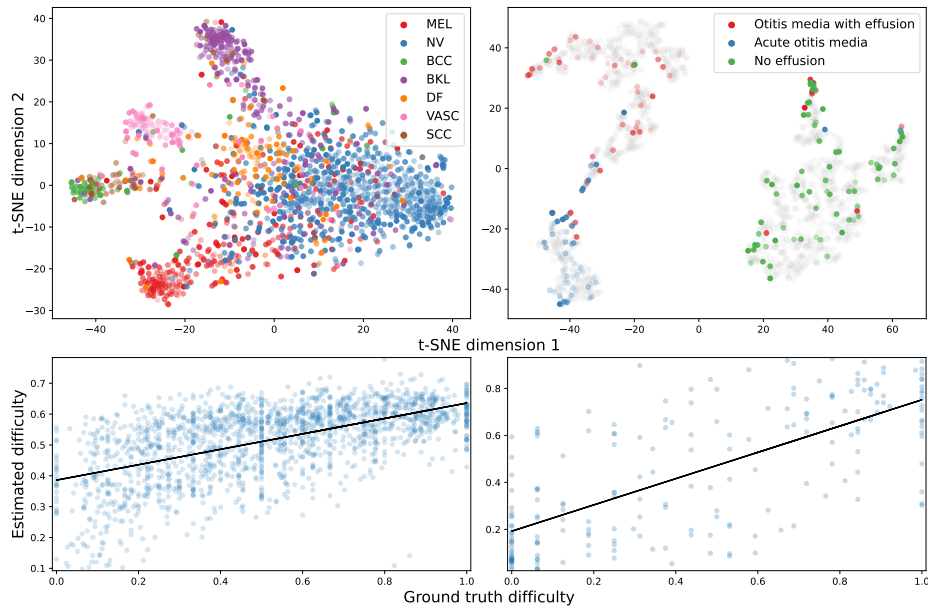
Fig. 2. Left: skin lesion dataset. **Right**: eardrum dataset. **Top**: visualization of the embeddings in two dimensions with t-SNE [17]. The transparency of each point indicates the ground truth difficulty with very transparent being the easiest. Grey points are the training samples for the eardrum data. **Bottom**: scatter plots of ground truth difficulties and difficulties estimated with the *embeddings + label* approach, together with the least squares regression lines.

are significantly better, with our best achieving Kendall's τ of 0.398 and 0.517. This corresponds to 69.9% and 75.8% of pairs being ordered correctly, which is an improvement of 12.8 and 11.1 percentage points from the best performing existing method (self-taught scoring).

Table 1 shows that our contribution of incorporating the ground truth class greatly increases performance. A similar tendency is seen in the higher performance of self-taught scoring compared to classification uncertainty, as the only difference between these two methods is the knowledge about the ground truth class. This intuition is also visible in Fig. 2, especially for the eardrum dataset, where the most difficult examples are often placed in the extremities of the clusters, or placed inside other clusters. This indicates that the embedding has a relation to difficulty, and shows the relevance of including the ground truth class label when estimating difficulty. Additionally, it is time-consuming and expensive to get multiple doctors to diagnose each case in a medical dataset, which is required to determine ground truth difficulties. Thus, the unsupervised methods proposed in this paper show great promise for estimating the difficulty directly from the image embeddings, without the need for ground truth difficulties.

Prospect of application: Our methods have demonstrated great potential in the estimation of human classification difficulty of medical images, which can be used to optimize and improve the training of medical professionals.

References

1. André, B., Vercauteren, T., Buchner, A.M., Shahid, M.W., Wallace, M.B., Ayache, N.: An image retrieval approach to setup difficulty levels in training systems for endomicroscopy diagnosis. In: MICCAI. pp. 480–487 (2010)
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML. pp. 41–48 (2009)
3. Brockhoff, P., Møller, J., Andersen, E., Bacher, P., Christiansen, L.: Introduction to statistics - enotes (2015)
4. Cha, D., Pae, C., Seong, S.B., Choi, J.Y., Park, H.J.: Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. EBioMedicine **45**, 606–614 (2019)
5. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
6. Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
7. Dagan, I., Engelson, S.P.: Committee-based sampling for training probabilistic classifiers. In: In Proceedings of the Twelfth International Conference on Machine Learning. pp. 150–157. Morgan Kaufmann (1995)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
9. Elio, R., Anderson, J.R.: The effects of information order and learning mode on schema abstraction. Memory & cognition **12**(1), 20–30 (1984)
10. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine learning **63**(1), 3–42 (2006)
11. Hacohen, G., Weinshall, D.: On the power of curriculum learning in training deep networks. In: ICML. pp. 2535–2544 (2019)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Kendall, M.G.: Rank correlation methods. (1948)
14. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: SIGIR'94. pp. 3–12. Springer (1994)
15. Li, X., Guo, Y.: Active learning with multi-label svm classification. In: Twenty-Third International Joint Conference on Artificial Intelligence (2013)
16. Liu, D., Xiong, Y., Pulli, K., Shapiro, L.: Estimating image segmentation difficulty. In: International Workshop on Machine Learning and Data Mining in Pattern Recognition. pp. 484–495. Springer (2011)
17. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

18. Musgrave, K., Belongie, S., Lim, S.N.: Pytorch metric learning (2020)
19. Nagle, F., Lavie, N.: Predicting human complexity perception of real-world scenes. Royal Society open science **7**(5), 191487 (2020)
20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
21. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Chen, X., Wang, X.: A survey of deep active learning. arXiv preprint arXiv:2009.00236 (2020)
22. Roads, B.D., Xu, B., Robinson, J.K., Tanaka, J.W.: The easy-to-hard training advantage with real-world medical images. Cognitive Research: Principles and Implications **3**(1), 1–13 (2018)
23. Sadasivan, V.S., Dasgupta, A.: Statistical measures for defining curriculum scoring function. arXiv preprint arXiv:2103.00147 (2021)
24. Sanakoyeu, A., Tschernezki, V., Buchler, U., Ommer, B.: Divide and conquer the embedding space for metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 471–480 (2019)
25. Scheidegger, F., Istrate, R., Mariani, G., Benini, L., Bekas, C., Malossi, C.: Efficient image dataset classification difficulty estimation for predicting deep-learning accuracy. The Visual Computer **37**(6), 1593–1610 (2021)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
27. Sundgaard, J.V., Harte, J., Bray, P., Laugesen, S., Kamide, Y., Tanaka, C., Paulsen, R.R., Christensen, A.N.: Deep metric learning for otitis media classification. Medical Image Analysis **71**, 102034 (2021)
28. Sundgaard, J.V., Värendh, M., Nordström, F., Kamide, Y., Tanaka, C., Harte, J., Paulsen, R.R., Christensen, A.N., Bray, P., Laugesen, S.: Inter-rater reliability of the diagnosis of otitis media based on otoscopic images and wideband tympanometry measurements. International Journal of Pediatric Otorhinolaryngology p. 111034 (2022)
29. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016)
30. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
31. Tudor Ionescu, R., Alexe, B., Leordeanu, M., Popescu, M., Papadopoulos, D.P., Ferrari, V.: How hard can it be? estimating the difficulty of visual search in an image. In: CVPR. pp. 2157–2166 (2016)
32. Tuia, D., Volpi, M., Copa, L., Kanevski, M., Munoz-Mari, J.: A survey of active learning algorithms for supervised remote sensing image classification. IEEE Journal of Selected Topics in Signal Processing **5**(3), 606–617 (2011)
33. Vijayanarasimhan, S., Grauman, K.: What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In: CVPR. pp. 2262–2269 (2009)
34. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: CVPR. pp. 5022–5030 (2019)
35. Wu, J., Sheng, V.S., Zhang, J., Li, H., Dadakova, T., Swisher, C.L., Cui, Z., Zhao, P.: Multi-label active learning algorithms for image classification: Overview and future promise. ACM Computing Surveys (CSUR) **53**(2), 1–35 (2020)