Modality-Classification of Microscopy Images Using Shallow Variants of Deep Networks

Juan Trelles Trabucco
Dept. Computer Science
University of Illinois at Chicago
Chicago, IL, USA
jtrell2@uic.edu

Pengyuan Li
Dept. Computer and Information Sciences
University of Delaware
Newark, DE, USA
pengyuan@udel.edu

Cecilia Arighi
Dept. Computer and Information Sciences
University of Delaware
Newark, DE, USA
arighi@dbi.udel.edu

Hagit Shatkay
Dept. Computer and Information Sciences
University of Delaware
Newark, DE, USA
shatkay@udel.edu

G. Elisabeta Marai
Dept. Computer Science
University of Illinois at Chicago
Chicago, IL, USA
gmarai@uic.edu

***Abstract*—Microscopy images are pervasive in biomedical research publications, where images obtained through various microscopy modalities (light, fluorescence, scanning, transmission) are often used to describe and summarize experiments and contributions. Hence, there is growing interest in automatically identifying these microscopy images' modality and utilizing this knowledge in automated search tools. However, identifying microscopy images poses challenges due to a lack of extensive collections of labeled images. We describe and evaluate two alternative approaches to microscopy image classification. In the first approach, we progressively fine-tuned layers of ResNet models. The second approach uses shallow variants of ResNet networks, where we leverage the outputs from previous convolutional blocks. We compare these results against a Support Vector Machine (SVM)-based baseline. Our results show that fine-tuning specific layers yields better results than fine-tuning the whole model. Furthermore, shallower variants produce competitive results when compared to the entire fine-tuned model.**

***Index Terms*—microscopy, image classification, deep learning**

I. Introduction

Microscopy images are pervasive in biomedical research papers. Researchers use microscopy images obtained through a variety of techniques to describe and summarize experiments and contributions. As such, there has been much interest in recent years to obtain and classify images from within publications [1]–[3].

Microscopes are ubiquitous in biomedical research; with the right setup, microscopy images reveal details otherwise hidden to the naked eye. Advanced microscopes come in a variety of configurations, offering different magnification powers and achievable resolutions. Yet, different types of microscopy are needed to address the wide spectrum of tasks (Fig. 1). For example, at a high level, in *light* microscopy, visible light passes through the sample and one or several lenses. A key advantage of this modality is enabling the inspection of living organisms, such as in rat surgery or cell analysis. A variation of light microscopy is *fluorescence* microscopy and its subtypes (confocal, photobleaching, etc.), where particular wavelengths excite fluorophores to enable the subsequent detection of the fluorescence signal. Last but not least, the electron microscopy family, with its subtypes *scanning* and *transmission*, allows researchers to obtain higher resolution images; however, these modalities are restricted to in vitro (non-living) samples.

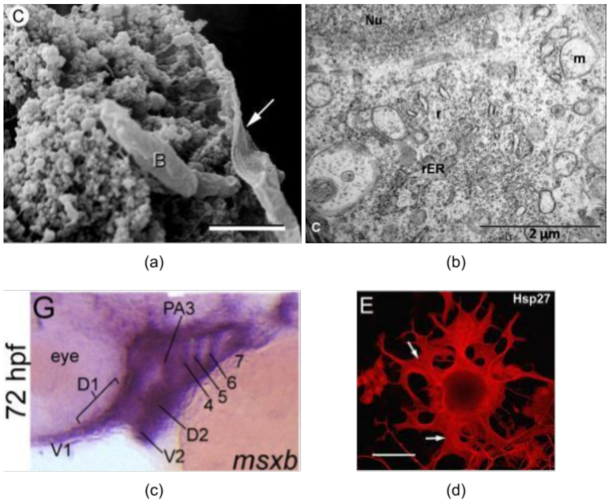
Classifying microscopy images by modality is a complicated task. Labels specifying the image modality subclass are seldom available. At a microscopic level, the imaging patterns themselves are difficult to distinguish, unlike in radiomics [4], [5]. Even for similar specimens, there are high inter-class and intra-class structural image variations. For example, light microscopy subclasses are similar to each other, as they all exhibit similar color features and patterns. Electron microscopy subclasses are also similar to each other, as they all exhibit grey scale features and similar structures. In addition, when producing a paper, the authors may edit the final illustrations. For example, the researchers may choose to publish colored-versions of gray-scale electron microscopy images in order to enhance their message. These edits can cause unexpected differences between images with similar image modalities and specimens.

Despite the paucity of labeled datasets, deep learning approaches obtain better classification results on the modality classification task than hand-crafted methods [6], [7]. Many of the deep learning models use pre-trained weights from ImageNet [8], which aims primarily natural photographs, despite the difference between nature and biomedical images. As a result, recent biomedical image classification work utilizes deep networks with transfer learning and fine tuning [9], [10]. This phenomenon has lead to an explosion in the size and number of parameters in the networks used for biomedical image classification. Whereas these large network solutions

This work has been supported by the U.S. National Institutes of Health through awards NLM-R01-LM012527 and NCI-R01-CA225190, and by the U.S. National Science Foundation through awards CNS-1828265 and CDSE-1854815.

report accuracy in the range of 76.87% to 88.48% [3], [10]–[13], those results focus on biomedical image datasets with up to 30 classes, where the classes are often considerably different from one another (e.g., X-Ray vs. microscopy). In contrast, the classification of microscopy image modalities within biomedical publications, where the challenge involves reduced labeled datasets and similar image content, is neither well-studied nor well-addressed.

In this paper, we propose and evaluate two alternative approaches to address microscopy image classification. Our work builds on a combination of Machine Learning techniques and thorough domain application. We investigate rigorously and empirically shallow variants of deep learners in this domain, along with the features to be used in classification. We analyze the conditions under which these shallow models are beneficial for modality classification tasks. The first approach follows the traditional approach of applying transfer learning and fine-tuning weights from pre-trained ResNet [14] models on ImageNet. For the second approach, we progressively remove the last convolutional block of ResNet models to create shallower variants. Our results highlight the importance of careful consideration of model depth when using small biomedical datasets that do not resemble general image repositories.



(a)

(b)

(c)

(d)

Fig. 1. Four microscopy images from the ImageCLEF dataset showcasing different modalities: (a) Electron microscopy (DMEL), (b) Transmission microscopy (DMTR), (c) Light microscopy (DMLI), and (d) Fluorescence microscopy (DMFL). As illustrated here, sample images retrieved from publications may also contain annotations, either as overlays or as border elements.

II. Methods

A. Data

We obtained our dataset by selecting the microscopy images from the ImageCLEF 2016 subfigure classification task [7]: light microscopy (DMLI), electron microscopy (DMEL), transmission microscopy (DMTR), and fluorescence microscopy (DMFL). Under this taxonomy, transmission microscopy is not part of the electron group, whereas electron microscopy encompasses scanning microscopy and other subclasses.

As the dataset is relatively small, we merged it with the microscopy images from the ImageCLEF 2013 dataset to obtain 2310 training images. We further performed data augmentation by random horizontal flips (p=0.5), and random rotations between 0 and 20 degrees. We chose the central 224x224 crop as our training sample. Besides, as the class distribution was highly unbalanced (Fig. 2), we oversampled from the DMEL and DMTR classes. To attain a similar distribution for our validation set, we used stratified sampling to gather 20% of the images. We further used for testing the ImageCLEF16 test set. The dataset encompasses figures extracted from biomedical literature; thus, the figures do not necessarily have the resolution of a raw microscopy image.

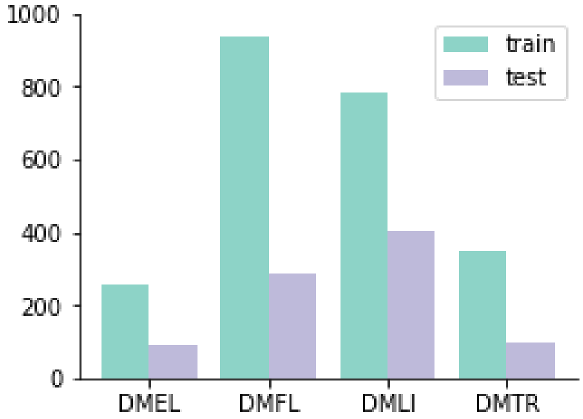


Fig. 2. Class distribution of the microscopy modality dataset. The dataset is unbalanced: fluorescence images (DMFL) are most common, followed by light microscopy images (DMLI). In contrast, the dataset features few samples of electron microscopy (DMEL) images and transmission microscopy (DMTR) images.

B. Transfer Learning, Fine-Tuning and Shallower Models

Compared to traditional machine learning approaches like Support Vector Machine (SVM), deep learning image classifiers rely on a considerable amount of training samples to train models with millions of parameters. Such a requirement is unfeasible in many domains; therefore, transfer learning and fine-tuning strategies provide alternatives to leverage existing trained models. In transfer learning, we reuse models pre-trained on a different dataset and retarget the model head (softmax layer) to match a new set of classes. We then train the new model on the target dataset; however, as only training the last layer may not yield the best performance, we can further fine-tune previous layers by updating their parameters. Two questions arise: what layers benefit from fine-tuning, and does a deeper network provide the most useful learned concepts?

To answer these questions, we experiment on the ResNet [14] family of models: ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152. At a high level, the ResNet architecture is a stack of convolutional blocks with skip connections called residual units that enable very deep models. This architecture organizes the residual units in four blocks (Fig. 3a); once starting a new block, each residual unit doubles the number of feature maps. Then, we characterize a base model, like ResNet50, by the number of residual units per block. For instance, ResNet50 has 3, 4, 6, and 3 residual units per block, respectively, while ResNet152 has the following configuration: 3, 8, 36, and 3. We chose the ResNet architecture for our experiments for their capability of creating deep models and for their competitive results in image classification.

For our first approach, we used transfer learning and fine-tuning for each ResNet variant. First, we updated the softmax layer to target our four microscopy classes. We created different variants by unfreezing the residual units' parameters from the deeper layers backward until we fine-tuned the whole model (Fig. 3b). For instance, the ResNet18 model has two residual units per model block, yielding eight variants (without including a fully-tuned model, and only updating the last fully connected layer). We trained every parameter from the softmax layer to the target residual for each variant, while the previous layer's parameters remained frozen. We used a constant learning rate of $5e-6$ for each fine-tuned model as it yielded more stable results than larger learning rates. We ended training 10 ResNet18 models, 18 ResNet34 models, 18 ResNet50 models, 35 ResNet101 models, and 52 ResNet152 models. Although ResNet34 and ResNet50 have the same number of residual units per block, from ResNet50, each residual block is more complex [1]; thus, there is an increase in the number of trainable layers.

Our second approach explored the effects of augmenting the number of residual units per block in the ResNet architecture. Although ResNet18 could be considered a shallower version of ResNet152, we shrank the architectures by removing the deeper layers one by one. As such, our shallower models did not yield the same configurations of the smaller base models. We pre-loaded the ImageNet weights for each trained model and fine-tuned the whole parameters with a constant learning rate of $5e-6$. Larger learning rates produced bumpier loss values during training.

In addition, we built an ensemble model by combining the outputs of the six deeper residual units of a ResNet50 model. We concatenated the outputs and fed them to a softmax layer (learning rate=$5e-6$). Our intuition was that we could leverage relevant features from different residual units, and potentially increase the classifier performance.

We trained each model for 100 epochs, and compared the model with the lowest validation loss value. We used a cross-entropy loss with an Adam optimizer. Additionally, we created a baseline based on a ResNet50 model pre-trained on ImageNet as a feature extractor, and an SVM linear classifier on top. We implemented our experiments using PyTorch and Scikit-learn [15], using our laboratory resources [16], [17]; our code and training reports [18] are available in this repository: github.com/jtrells/biomedical-image-classification[2].

III. Results

A. Fine-tuned models

Results of the fine-tuned models (Table I) show that a fine-tuned ResNet18 obtained the highest test accuracy (89.00). Compared to other ResNet18 models, the best performing variant (*block 1-1*) was marginally better. Notably, this model was 3.2 percentual points better than only fine-tuning the softmax layer. When comparing the deeper models, we found that the best ResNet50 and ResNet152 models obtained the same accuracy. Notably, for each base model, fine-tuning the whole model did not yield the highest accuracy. Still, on average, these full models were only 0.87 percentual points worse than the best performing variant.

Coefficients of the Pearson correlation between the number of trainable parameters and test accuracy show decreasing correlation as a model gets bigger. Smaller models like ResNet18 and ResNet35 have a strong correlation ($r = 0.81$, $r = 0.75$ respectively), the middle-sized ResNet50 has a moderate correlation ($r = 0.56$), and the large ResNet101 and ResNet152 have a moderate ($r = 0.64$) and weak correlation ($r = 0.27$).

Arguably, fine-tuning a full ResNet model is more economical than looking for the specific layer that provides better performance. To this end, Fig. 4 shows the variability in the accuracy results, where for each line chart the number of trainable parameters increases from left to right.

TABLE I
Summary statistics for fine-tuned models group by base model. Last column shows the accuracy difference between the best performing model (block in brackets) and a fully-tuned model.

Model	avg	std	max	min	diff (best-full)
ResNet18	88.25	1.16	**89.00**	84.88	0.92 [b1-1]
ResNet34	86.86	0.31	87.29	86.03	0.57 [b3-2]
ResNet50	87.44	0.97	88.89	84.42	1.15 [b3-0]
ResNet101	88.11	0.54	88.77	86.14	0.69 [b3-3]
ResNet152	88.06	0.75	88.89	84.77	1.03 [b3-35 to b3-32, b1-2]

B. Shallower models

With respect to shallower models, the best performing models were indeed a shallower version of a ResNet base model. ResNet152 *block 3-11* (*block 3* has 36 residual units) model yielded the highest accuracy, closely followed by the ResNet101 *block 3-18* and the ResNet34 *block 3-0* models. Compared to the fine-tuned model evaluation, the difference between the best variant and the full model was higher, with an average of 1.70 percentual points (Table II).

Shallowest variants led to the worst accuracy scores among all base models. Yet, the shallowest models from ResNet50, ResNet101, and ResNet152 are significantly worst

[1] Bottleneck module on torchivision ResNet model

[2] GradCAM++ [19] implementation from the GitHub repository 1Konny/gradcam_plus_plus-pytorch

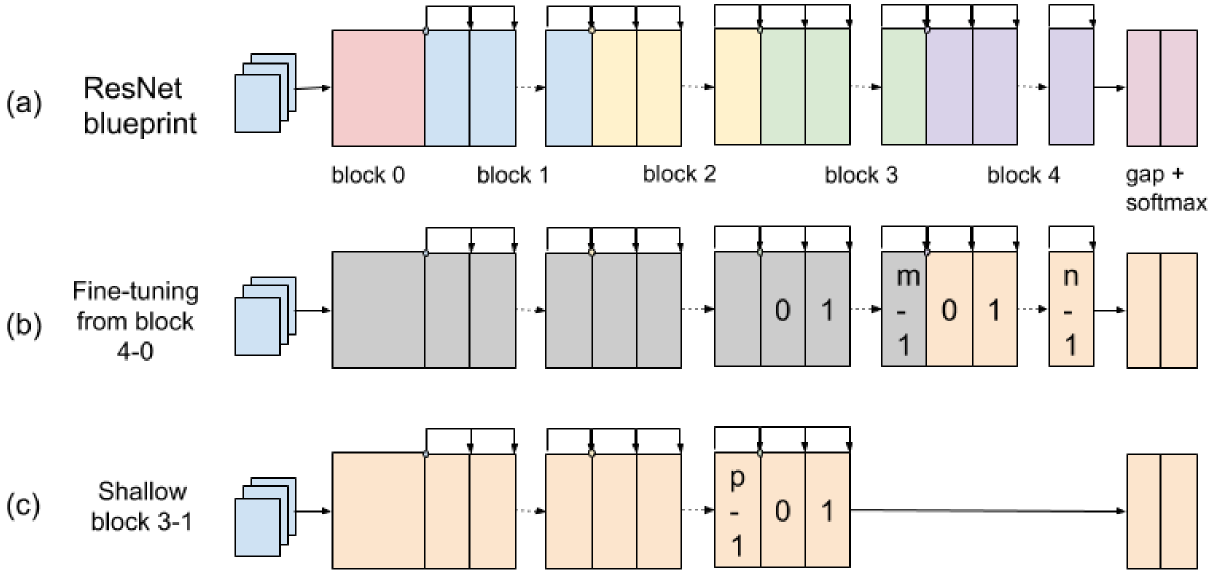


Fig. 3. (a) ResNet architecture blueprint. The number of residual blocks varies per ResNet base model. (b) Fine-tuning a base model from up to the first residual unit of block 4. (c) Shallow model using the outputs from block 3-1. Grey boxes represent layers with frozen parameters; orange boxes represent fine-tuned layers.

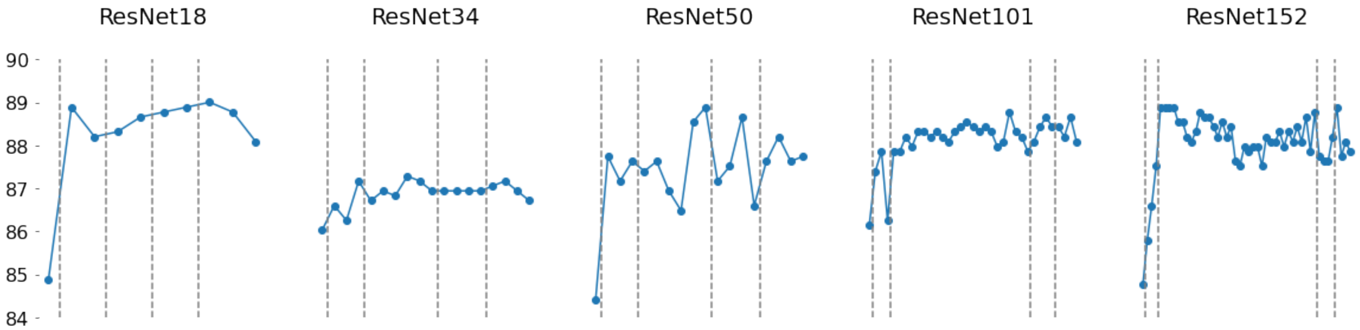


Fig. 4. Test accuracy for fine-tuned variants (y-axis limit between 84 and 90 percent). For each ResNet base model, the number of trainable parameters per variant **increases** from left to right. Vertical lines divide the architectural blocks: softmax layer, block 4, block 3, block 2, and block 1 + fully-tuned model.

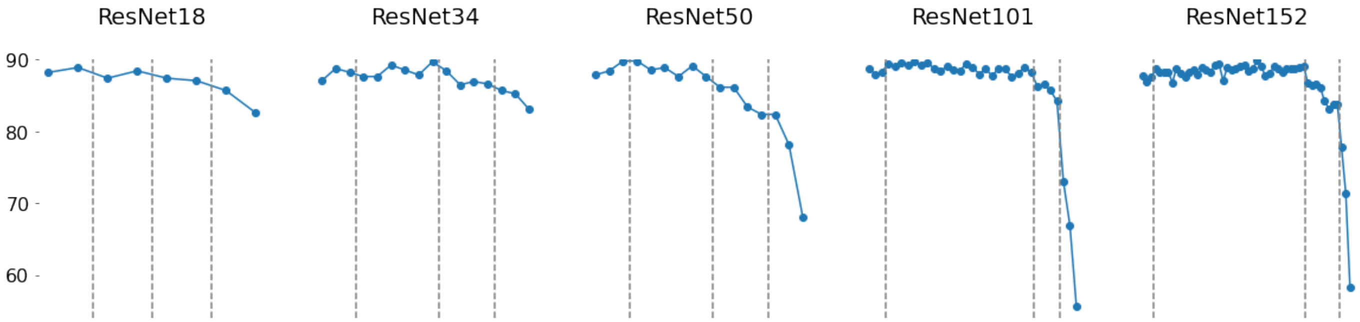


Fig. 5. Test accuracy for shallow variants (y-axis limit between 54 and 90 percent). For each ResNet base model, the number of trainable parameters per variant **decreases** from left to right. Vertical lines divide the architectural blocks: block 4, block 3, block 2, and block 1. We note that the shallowest models yielded the worst accuracy.

than ResNet18 and ResNet34 (Fig. 5). After reaching *block 3* in the architecture, we start to obtain good accuracy results, then a bottleneck becomes apparent (especially for ResNet101 and ResNet152).

TABLE II
SUMMARY STATISTICS FOR SHALLOWER RESNET MODELS GROUP, BY BASE MODEL. LAST COLUMN SHOWS THE ACCURACY DIFFERENCE BETWEEN THE BEST PERFORMING MODEL (IN BRACKETS) AND THE FULL MODEL.

Model	**avg**	**std**	**max**	**min**	**diff (best-full)**
ResNet18	86.96	1.89	88.89	82.59	0.69 [b4-0]
ResNet34	87.31	1.62	89.69	83.05	2.63 [b3-0]
ResNet50	85.27	5.46	88.80	68.04	1.95 [b4-0, b3-5]
ResNet101	86.21	7.10	89.69	58.19	1.03 [b3-18]
ResNet152	86.76	5.10	**89.92**	58.19	2.18 [b3-11]

C. Comparison with baseline

In Table III, we show the comparison of the SVM baseline, our best fine-tuned and shallow models, our ensemble of shallow ResNet50 variants, and two related models. We chose these two last models [10], [13] for our comparison as the authors provided precision, recall, and F1 scores per class, although the authors trained on the 30 classes from ImageCLEF instead of specifically the four microscopy classes. These scores provide a better baseline than the SVM model that uses extracted features from ImageNet.

Our results show that our models performed better than the baseline. Our best accuracy score over this dataset was 89.92%. In general, we obtained lower scores with the electron microscopy class (DMEL), for which we had the lowest number of training samples.

D. Wrong predictions

Despite having trained a wide variety of models during both approaches, we found that all models failed to correctly predict a specific set of testing images. For instance, Fig. 6 shows an electron microscopy figure incorrectly classified as a fluorescence image by the best ResNet50 fine-tuned variant. The original figure, on the top left, presents an unusual red and green coloring. To its right, we show the gradient activations of two layers using GradCAM++ [19]. Activations on *block 3-5* suggest a higher interest in the colored nuclei. Changing the figure colormap to grayscale (bottom left) yields to a correct prediction, and the interpretability approach indicates that there is less focus on the colored elements. Yet, it is not clear if the network focus was centered on color or textures.

We found that the shallowest models correctly predicted some figures that their bigger counterparts failed to predict. However, these models had the lowest training and test accuracy; an inspection of the training accuracy and loss charts suggests that the models lack capacity, and thereby their predictions are not trustworthy.

IV. DISCUSSION

Our results indicate that, as expected, fine-tuning deep residual units on the microscopy dataset while keeping the initial parameters frozen yields better results. Yet, finding the right fine-tuning configuration is a tedious task that may only worth the effort for small datasets (we spent between 15 and 30 minutes training each model). Compared to fine-tuning the whole model, finding the best performing models yielded a gain of 0.86 percentual points on average. Therefore, unless that difference is a considerable gain, fine-tuning the whole model is the recommended approach. Another interesting avenue to explore on our dataset is the adaptive transfer learning strategy where each layer is trained depending on the input sample [20].

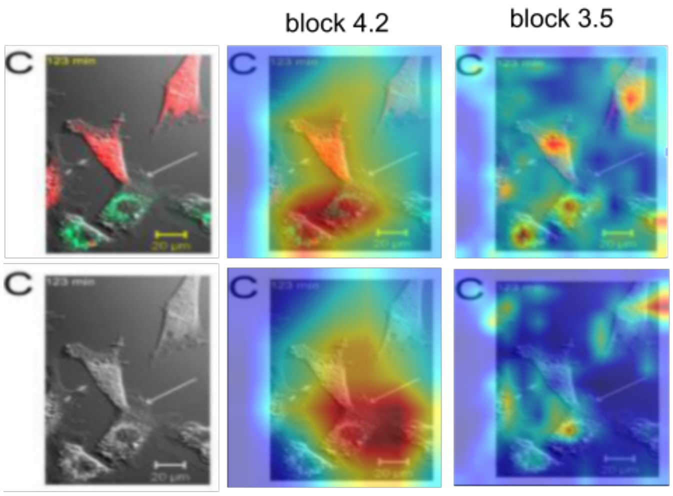


Fig. 6. Top left: Microscopy electron image incorrectly classified as a fluorescence microscopy image. Bottom left: The same image in grayscale; now, the classifier correctly predicted the image class. GradCam++ activation maps for two different layers are shown to the right.

Given our dataset's size, it is not surprising that the fine-tuned ResNet18 model obtained competitive performance compared to its deeper counterparts. ResNet18 *block 1-1* was only 0.92% worse than the best shallow ResNet152 *block 3-11*, but we spent much less time finding this model configuration (ResNet152 had 52 different shallower variants). The ResNet152 results showcase the effectiveness of residual units.

Results from our second approach using shallow models indicate that extracted features from previous residual units contain relevant information. In part, we believe that this is also related to our dataset's small size; while shrinking its capacity, the model is less prone to overfitting and can still learn useful features. Fig. 5 suggests that the most valuable features mostly appear from *block 3*, but it is still difficult to determine the location quickly.

We believe that identifying previous layers with useful features allows us to expand our classifier hierarchically. For example, microscopy classes have experimental methods that depend on the modality (e.g., In Situ Hybridization for light microscopy). A hierarchical classifier can obtain information from a previous layer to identify the modality while the deeper layers learn features specialized in the methods.

TABLE III
PERFORMANCE METRICS FROM THE BASELINE (PRE-TRAINED MODEL ON IMAGENET + SVM), OUR BEST PERFORMING MODELS AND RELATED WORK.

Models	**Accuracy**	**Metrics**	**DMEL**	**DMFL**	**DMLI**	**DMTR**
Baseline: ResNet50 (pre-trained on ImageNet) + SVM	85.80	*Precision*	54.51	91.07	90.74	67.07
		Recall	42.05	93.31	96.79	57.29
		F1	47.44	92.17	93.67	61.80
Best Fine-Tuned: ResNet18 *block 1-1*	89.00	*Precision*	61.90	93.10	93.13	76.53
		Recall	44.32	**95.07**	97.04	78.12
		F1	51.66	**94.08**	95.04	77.32
Best Shallow: ResNet152 *block 3-11*	89.92	*Precision*	**72.88**	**94.33**	92.31	**77.67**
		Recall	**48.86**	93.66	97.78	**83.33**
		F1	**58.50**	93.99	94.96	80.40
Shallow ResNet50 Ensemble	89.58	*Precision*	64.52	93.71	**93.63**	76.24
		Recall	45.45	94.37	**98.02**	80.21
		F1	53.33	94.04	**95.78**	**78.17**
InceptionV4 Ensemble (*Koitka and Friedrich* [10])	–	*Precision*	62.96	82.21	91.63	73.17
		Recall	38.64	94.37	91.85	62.50
		F1	47.89	87.87	91.74	67.42
AlexNet + GoogleLeNet Ensemble (*Kumar et al.* [13])	–	*Precision*	31.46	91.97	86.93	48.38
		Recall	31.82	88.73	80.49	62.50
		F1	31.64	90.32	83.59	54.55

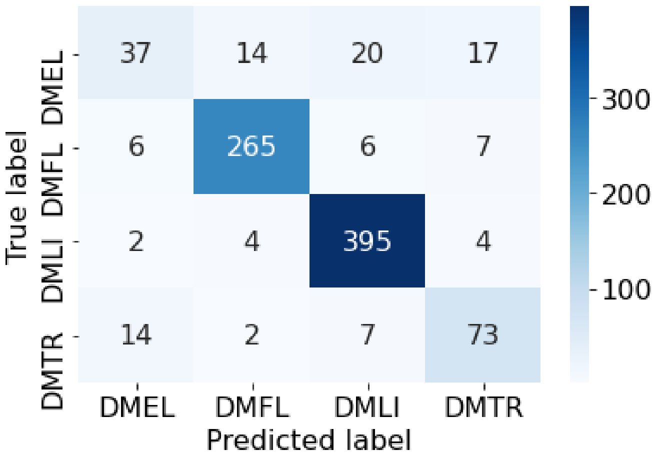


Fig. 7. Confusion matrix for model ResNet152 *block 3-11*

Our best performer model analysis confirms that specific microscopy classes are harder to classify (Fig. 7). Most confusion happened between the DMTR (transmission) and DMEL (electron) microscopy classes. One explanation is the similar color distribution (mostly gray-scale images), compared to the abundance of color in DMLI (light) and DMFL (fluorescence) microscopy classes. The second set of misclassifications happened between DMEL against DMLI. We suspect that the main issue is that post-processed DMEL images for camera-ready papers use colors initially not present in the original DMEL class distribution (Fig. 6). Although we tried to diminish the unbalanced dataset's effect, the two classes with fewer samples (DMEL and DMTR) constrain our classifier's effectiveness.

In terms of limitations, our approach does not identify the most recommended shallow variant without an extensive search. In particular, we cannot conclude that adding more layers (and trainable parameters) correlates linearly with the model accuracy. A possible explanation is the ResNet residual design, where the network connections do not strongly depend on each other [21]. Consequently, removing layers from the ResNet architecture does not necessarily degrade the network performance as it happens with stacking architectures like VGG [22]. Besides, although many biomedical classification tasks suffer from the unbalanced dataset problem [23], we cannot generalize our findings without further tests.

V. RELATED WORK

The ImageCLEF subfigure classification by image modality task boosted the development of modality classification approaches. Since the task introduction, traditional methods have focused on feature engineering [9], and deep learning approaches have been used either in isolation or as part of ensemble models [7]. For example, Koitka and Friedrich [9] used a ResNet-152 model with transfer learning and an ensemble of SVM and a ResNet-152 model. Kumar et al. [13] also proposed an ensemble model of a fine-tuned AlexNet [24] model, a fine-tuned InceptionV1 [25] model, and three one-vs-one SVM models leveraging the outputs from the two network models.

Notably, deep learning approaches dominate the modality classification task [10]. Koitka and Friedrich [10] used an ensemble model of fine-tuned InceptionV4 [26] models. Yu et al. [11] evaluated an ensemble model of a VGG16 [22], a ResNet-50 [14] fine-tuned model, and a six-layered CNN model trained from scratch. Andrearczyk and Muller [3] used a DenseNet-169 [27] model to perform multimodal training using figures and captions. Finally, Zhang et al [12] leveraged CNN features to feed a synergic network. Previous work, however, leverages only the outputs from the last layers of convolutional neural networks. Therefore, we focus on previous layers to evaluate their benefits on an unbalanced and small dataset.

VI. Conclusion

In this work, we described and evaluated two alternative approaches to microscopy image classification. The first approach is based on progressively fine-tuning earlier layers of ResNet networks. The second approach focuses on fine-tuning shallower versions of ResNet models. When fine-tuning ResNet variants, we found that a ResNet18 variant yielded the highest accuracy. Also, we believe that fine-tuning the whole model can be a good enough approach rather than finding the right combination of fine-tuned layers.

In contrast, for shallow ResNet variants, we found that a ResNet152 variant obtained the best performance. For shallow variants, features from the third block produced higher accuracy scores. Finally, the best shallow model outperformed the best-fine-tuned variant. Still, given the size of our dataset, the margin was relatively small compared to the cost of training multiple models.

Acknowledgment

We thank all members of the Electronic Visualization Laboratory, and in particular Lance Long and Andrew Wentzel, for their feedback and technical support. This work has been supported by the U.S. National Institutes of Health through awards NLM-R01-LM012527 and NCI-R01-CA225190, and by the U.S. National Science Foundation through awards CNS-1828265 and CDSE-1854815.

References

[1] D. Kim, B. P. Ramesh, and H. Yu, "Automatic figure classification in bioscience literature," *Journal Biomedical Informatics*, vol. 44, no. 5, pp. pp. 848–858, 2011.

[2] T. Kuhn, M. L. Nagy, T. Luong, and M. Krauthammer, "Mining images in biomedical publications: Detection and analysis of gel diagrams," *Journal Biomedical Semantics*, vol. 5, no. 1, p. pp. 10, 2014.

[3] V. Andrearczyk and H. Müller, "Deep multimodal classification of image types in biomedical journal figures," in *Proceedings of International Conference Cross-Language Evaluation Forum for European Languages*. Springer, 2018, pp. 3–14.

[4] H. Elhalawani, T. A. Lin, S. Volpe, A. S. Mohamed, A. L. White, J. Zafereo, A. J. Wong, J. E. Berends, S. AboHashem, B. Williams *et al.*, "Machine learning applications in head and neck radiation oncology: lessons from open-source radiomics challenges," *Frontiers in oncology*, vol. 8, p. 294, 2018.

[5] H. Elhalawani, A. S. Mohamed, A. L. White, J. Zafereo, A. J. Wong, J. E. Berends, S. AboHashem, B. Williams, J. M. Aymard, A. Kanwar *et al.*, "Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges," *Scientific data*, vol. 4, p. 170077, 2017.

[6] O. Pelka and C. M. Friedrich, "FHDO Biomedical Computer Science Group at medical classification task of ImageCLEF 2015," in *Working Notes of CLEF*, vol. 1391, 2015.

[7] A. García Seco de Herrera, R. Schaer, S. Bromuri, and H. Müller, "Overview of the ImageCLEF 2016 medical task," in *Working Notes of CLEF*, 2016.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[9] S. Koitka and C. M. Friedrich, "Traditional feature engineering and deep learning approaches at medical classification task of imageclef 2016." in *Working Notes of CLEF*, 2016, pp. 304–317.

[10] S. Koitka and C. Friedrich, "Optimized convolutional neural network ensembles for medical subfigure classification," in *Proceedings of International Conference Cross-Language Evaluation Forum for European Languages*. Springer, 2017, pp. 57–68.

[11] Y. Yu, H. Lin, J. Meng, X. Wei, H. Guo, and Z. Zhao, "Deep transfer learning for modality classification of medical images," *Information*, vol. 8, no. 3, p. pp. 91, 2017.

[12] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Medical image classification using synergic deep learning," *Medical Image Analysis*, vol. 54, pp. pp. 10–19, 2019.

[13] A. Kumar, J. Kim, D. Lyndon, M. Fulham, and D. Feng, "An ensemble of fine-tuned convolutional neural networks for medical image classification," *IEEE Biomedical and Health Informatics*, vol. 21, no. 1, pp. pp. 31–40, 2016.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] M. Monfort, T. Luciani, J. Komperda, B. Ziebart, F. Mashayek, and G. E. Marai, "A deep learning approach to identifying shock locations in turbulent combustion tensor fields," in *Modeling, Analysis, and Visualization of Anisotropy*. Springer, 2017, pp. 375–392.

[16] G. E. Marai, A. G. Forbes, and A. Johnson, "Interdisciplinary immersive analytics at the electronic visualization laboratory: Lessons learned and upcoming challenges," in *2016 Workshop on Immersive Analytics (IA)*. IEEE, 2016, pp. 54–59.

[17] G. E. Marai, J. Leigh, and A. Johnson, "Immersive analytics lessons from the electronic visualization laboratory: a 25-year perspective," *IEEE computer graphics and applications*, vol. 39, no. 3, pp. 54–66, 2019.

[18] L. Biewald, "Experiment tracking with weights and biases," 2020, software available from wandb.com. [Online]. Available: https://www.wandb.com/

[19] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839–847.

[20] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, "Spottune: transfer learning through adaptive fine-tuning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4805–4814.

[21] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Adv. Neural Information Processing Systems*, 2016, pp. 550–558.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *preprint arXiv:1409.1556*, 2014.

[23] C. Zhang, W. Tavanapong, J. Wong, P. C. de Groen, and J. Oh, "Real data augmentation for medical image classification," in *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, 2017, pp. 67–76.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *preprint arXiv:1602.07261*, 2016.

[27] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.