ResNet18 Performance: Impact of Network Depth and Image Resolution on Image Classification

Haixia Liu
Computer Science and Creative Technologies
UWE
Bristol, UK
haixia.liu@uwe.ac.uk

Tim Brailsford
Computer Science and Creative Technologies
UWE
Bristol, UK
Tim.Brailsford@uwe.ac.uk

Larry Bull
Computer Science and Creative Technologies
UWE
Bristol, UK
Larry.Bull@uwe.ac.uk

Abstract

This study explores aspects of using ResNet18 for the classification of biomedical images, using the 2D medical images from the MedMNIST library that is widely used for benchmarking. The number of layers within the ResNet18 was studied. The typically used depth (block 4) was generally found to be robust across a variety of datasets at the recommended image resolution (224x224). We found that decreasing the resolution while maintaining block 4 depth can significantly improve performance in some, but not all, cases. The effects of varying both depth and resolution simultaneously were evaluated, and we found a non-linear relationship between depth/resolution and performance. The context of the images seems to be important, with the best performing combinations of network depth and image resolutions varying. We examined the feature maps, and found them to be very variable for the best performing models.

CCS Concepts

• **Computing methodologies** → **Neural networks**; • **Computing methodologies** → **Computer vision**; • **Applied computing** → **Life and medical sciences**;

Keywords

ResNet18, Deep Learning, Image Classification, Explainable AI (XAI)

ACM Reference Format:
Haixia Liu, Tim Brailsford, and Larry Bull. 2024. ResNet18 Performance: Impact of Network Depth and Image Resolution on Image Classification. In *2024 The 8th International Conference on Advances in Artificial Intelligence (ICAAI 2024), October 17–19, 2024, London, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3704137.3704173

1 Introduction

The usual approach to image classification using ResNet18 [5] and ImageNet [2] involves resizing images to 224x224 before processing with a convolutional neural network (CNN). The prevailing view is that this is a good compromise between the amount of information in the image and the computational requirements of training [4]. Despite the widespread use of this size, there is little consensus that it always produces the best results and it is quite likely that the context of use is important. It is usually assumed that deeper networks perform better than shallow ones, and Shamir [11] has provided a theoretical justification as to why this is.

The debate about the relative effectiveness of deep versus shallow learning is ongoing, with various studies offering different perspectives. For example, Robles Herrera et al. [9] have suggested that deep learning is more effective in situations with natural symmetries, while shallow learning may be better in other cases. Malach and Shalev-Shwartz [6] have explored the relationship between the expressivity of deep networks and their training efficiency, finding that the efficacy of deep networks depends on whether the distribution can be well approximated by shallower networks. Mhaskar et al. [7] have provided a theoretical basis for the effectiveness of deep convolutional networks in approximating compositional functions with fewer parameters. Furukawa and Zhao [3] have focused on the interpretability of deep learning, finding that simpler and better rules can be extracted from higher layers of deep networks. Picon Ruiz et al. [8] have provided a broader perspective, explaining that deep learning often outperforms more traditional machine learning due to its ability to handle complex pattern recognition tasks. However, Cirrincione et al. [1] have found that for a gear fault diagnosis task, shallow networks generally outperform deep ones. Recent studies by Zhang et al. [13] have incorporated dendritic learning into convolutional neural networks, in an attempt to more closely mimic biological neurons' information processing capabilities. They have used a shallow architecture of pyramidal neurons for image classification. Despite all of this, the relationships between image size, network depth and performance are still not well understood, especially in medical images, rather than the more commonly studied ImageNet images [2].

In this paper, the research questions that we have been addressing are:

- For each MedMNIST dataset explored in this study, are there better, shallower depths rather than 4 blocks based on 224x224 resolution of the standard version of ResNet18?
- For each MedMNIST dataset explored in this study, are there better image resolutions in comparison to 224x224 based on a depth of 4 blocks?
- For each MedMNIST dataset explored in this study, are there better combinations of depths and resolution rather than using the 4 blocks and 224x224 resolution?

ICAAI 2024, October 17–19, 2024, London, United Kingdom

ACM ISBN 979-8-4007-1801-4/24/10
https://doi.org/10.1145/3704137.3704173

- What are the feature map sizes of the final CNN layer in the best performing models? Do they maintain the same size, or do they vary? If they differ, what might be the underlying reasons?

2 Methodology

2.1 Datasets

Several 2D medical image datasets (MedMNIST) [12] were used in this study, which have been widely used for example in Zhang et al. [13]. Table 1 shows the details of these datasets, with sample size indicating the total number of samples in each. These datasets vary in size and number of classes, with up to 58,850 samples and 11 classes. Images were pre-processed utilizing several techniques including resizing, normalisation.

Image resizing: The 2D medical image data were obtained from the MedMNIST library through its API [1]. The images used were already of the sizes that were used in the original study [12] (e.g.: 28x28, 64x64, 128x128 and 224x224).

Normalization: Input values were normalized using the transforms library[2], with mean and standard deviation parameters set to 0.5. This process involves subtracting the mean and dividing by the standard deviation for each channel, which is one of the basic pixel value scaling techniques [10].

Training: We used *stratify=y* during data splitting to ensure that the class distribution is maintained in the training, validation and testing sets. Sixty percent of the data was used for training, twenty percent for validation and twenty percent for testing.

Table 1: The datasets used in this study. The image sizes for each dataset include 28x28, 64x64, 128x128 and 224x224.

Dataset name	Sample size	Image Type
BloodMNIST	17,092 (8 classes)	color
BreastMNIST	780 (Binary-Class)	gray
DermaMNIST	10,015 (7 classes)	color
PneumoniaMNIST	5,856 (Binary-Class)	gray
RetinaMNIST	1,600 (5 classes)	color
OrganAMNIST	58,850 (11 classes)	gray
OrganCMNIST	23,660 (11 classes)	gray
OrganSMNIST	25,221 (11 classes)	gray

2.2 Feature Map Size

The fundamental unit utilised for building ResNet18 is called a "basic block", which is a concept introduced in the original ResNet18 paper [5] , and illustrated in Table 3.

Unlike many other studies that go deeper using compound scaling, we go shallower and down scale the image resolution, in an attempt to achieve similar results using less computational power. Table 2 shows the feature maps size per block per image resolution.

[1]https://medmnist.com/
[2]https://pytorch.org/vision/stable/transforms.html

Table 2: Feature map sizes for each block for different image resolutions.

Resolution	Block1	Block2	Block3	Block4
28x28	7x7	4x4	2x2	1x1
64x64	16x16	8x8	4x4	2x2
128x128	32x32	16x16	8x8	4x4
224x224	56x56	28x28	14x14	7x7

Table 3: Description of the basic block.

Layer	Description
conv1	Convolutional layer with kernel size 3x3, padding of 1
bn1	Batch normalization layer
conv2	Convolutional layer with kernel size 3x3, padding of 1
bn2	Batch normalization layer
relu	Rectified Linear Unit (ReLU) activation function
identity_downsample	Optional downsampling function

2.3 Experiment Settings and Design

The experiments were implemented in Python using Pytorch[3]. During training, all the experiments were done under the same settings: the maximum number of epochs was set to 50; batchsize was 32; Stochastic Gradient Descent (SGD) optimizer with a learning rate 0.001 and momentum 0.9 were used. The early stopping strategy was used with a patience of 10, and cross entropy loss [4] was monitored on the validation set. To perform statistical tests (students ttest), each model was run with 10 different randomstates [5] which were consistent among all experiments, to enable like for like comparisons. The network was trained from scratch without using any pre-trained models, and no augmentation techniques were applied.

3 Results

All the results reported and discussed are on testing sets based on the average of 10 runs with different randomstates.

3.1 Varying Depth with Standard Image Resolution

Figure 1 shows how using shallower versions of the standard ResNet18 architecture affects the classification accuracy on the benchmark datasets used here. Performance generally improves with increasing depth, e.g., OrganSMNIST shows the typical behaviour often reported in the literature. However, other depths appear at least competitive on some datasets, e.g., DermaMNIST, and as shown in Table 4, in one case a slightly shallower CNN outperformed the standard ResNet.

[3]https://pytorch.org/
[4]https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html
[5]https://numpy.org/doc/1.14/reference/generated/numpy.random.RandomState.html

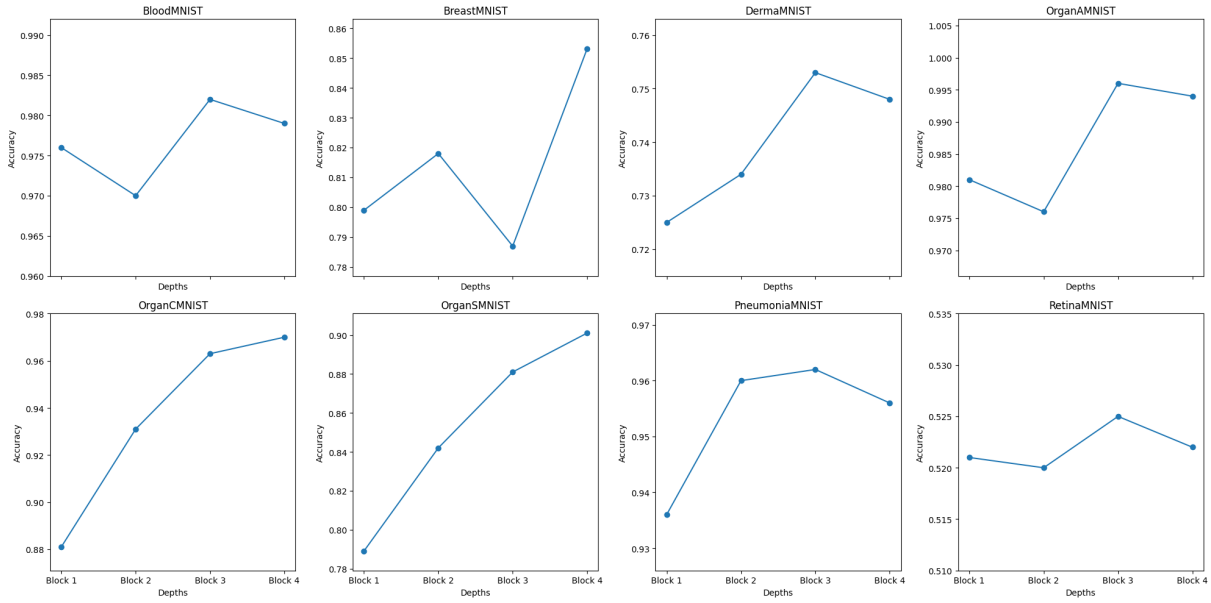

Figure 1: Each plot represents the accuracy of models evaluated at different blocks (block 1 through block 4) for all datasets. To visualize the changes the y-axis range has been adjusted to provide a clearer view of the variations.

Table 4: The result of exploring depths for all the datasets with 224x224 that are statistically significant.

Dataset name	Best Depth	Significance
OrganAMNIST	Block3	p<0.05 *

3.2 Varying Image Resolution with Standard Depth

Figure 2 shows how using lower image resolutions than the recommended ResNet18 approach affects the classification accuracy on the benchmark datasets used here. Performance often improves or is highest with increasing resolution, e.g., BloodMNIST. However, other resolutions appear at least competitive on some datasets, e.g., PneumoniaMNIST, and as shown in Table 5, in a number of cases a lower image resolution significantly outperforms the recommended standard of 224x224.

Table 5: The results of exploring resolutions for all datasets with the depth being block 4, showing only those that are statistically significant.

Dataset name	Resolution	Significance
PneumoniaMNIST	64x64	p<0.01 **
	128x128	p<0.05 *
OrganAMNIST	64x64	p<0.05 *
OrganSMNIST	64x64	p<0.001 ***
	128x128	p<0.001 ***

3.3 Varying Both Image Resolution and Depth

These results show that performance can be significantly improved by decreasing either depth or image resolution from the standard ResNet18. Figure 3 shows how varying both simultaneously can affect the classification accuracy across a wide range of the parameter space on the benchmark datasets. Table 6 details all of the combinations that significantly outperform the recommended standard depth of 4 blocks with 224x224 resolution. The two fundamental aspects of ResNet18, and by implication CNN in general, behave in non-linear ways. The detailed average accuracy across 10 runs for each dataset under each condition is shown in Table 7.

4 Conclusion and Discussion

We have explored two approaches of using neural networks for the classification of biomedical images. We have adopted the commonly used ResNet18 as our baseline and compared it with shallower models using lower-resolution images. The results suggest that the standard depth is robust when using image sizes of 224x224. However, this is not necessarily the case using lower resolutions. Results obtained by simultaneously varying depth with resolution suggest a non-linear behaviour following one of three patterns. The expected one is that performance often increases with depth regardless of resolution (e.g. Figure 3 using OrganCMNIST). However, this is not always the case and sometimes increasing depth and/or resolution provides no significant benefit in classification performance (e.g. Figure 3 using BloodMNIST). In some other examples the performance is sensitive to both the resolution and depth combination (e.g. Figure 3 using PneumoniaMNIST). For the best-performing combinations, the feature maps of the top models vary across different datasets and even within the same dataset (e.g., in Table 6 for OrganAMNIST, the feature map sizes differ significantly, with

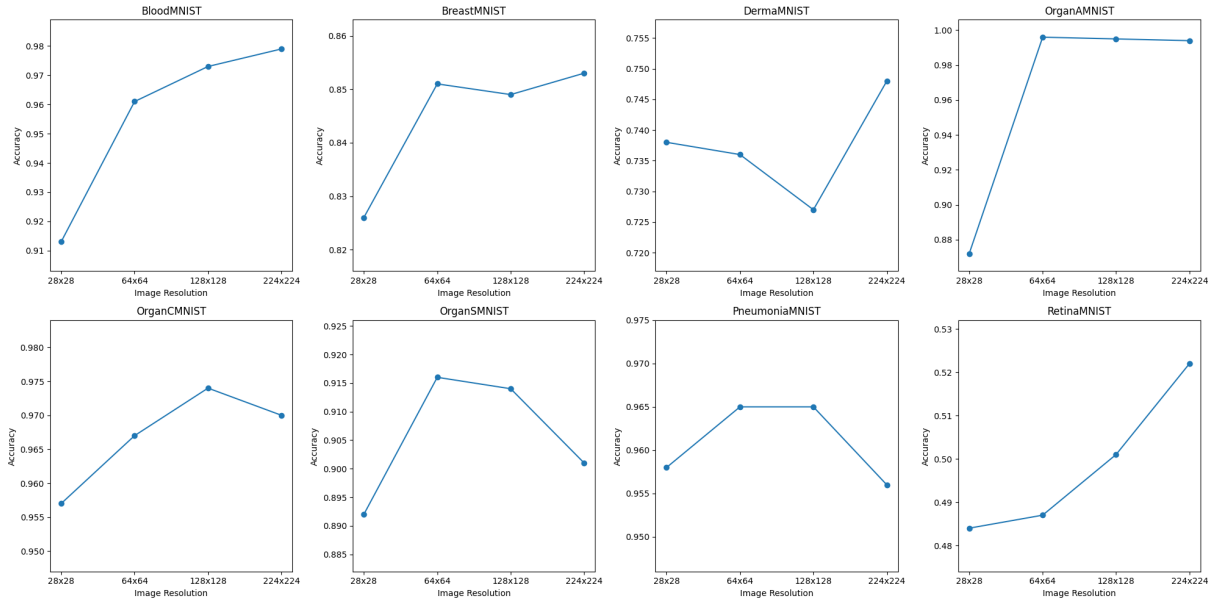

Figure 2: The accuracy of models evaluated at block 4 across different image resolutions, ranging from 28x28 to 224x224. The y-axis range has been scaled to emphasize the differences in accuracy.

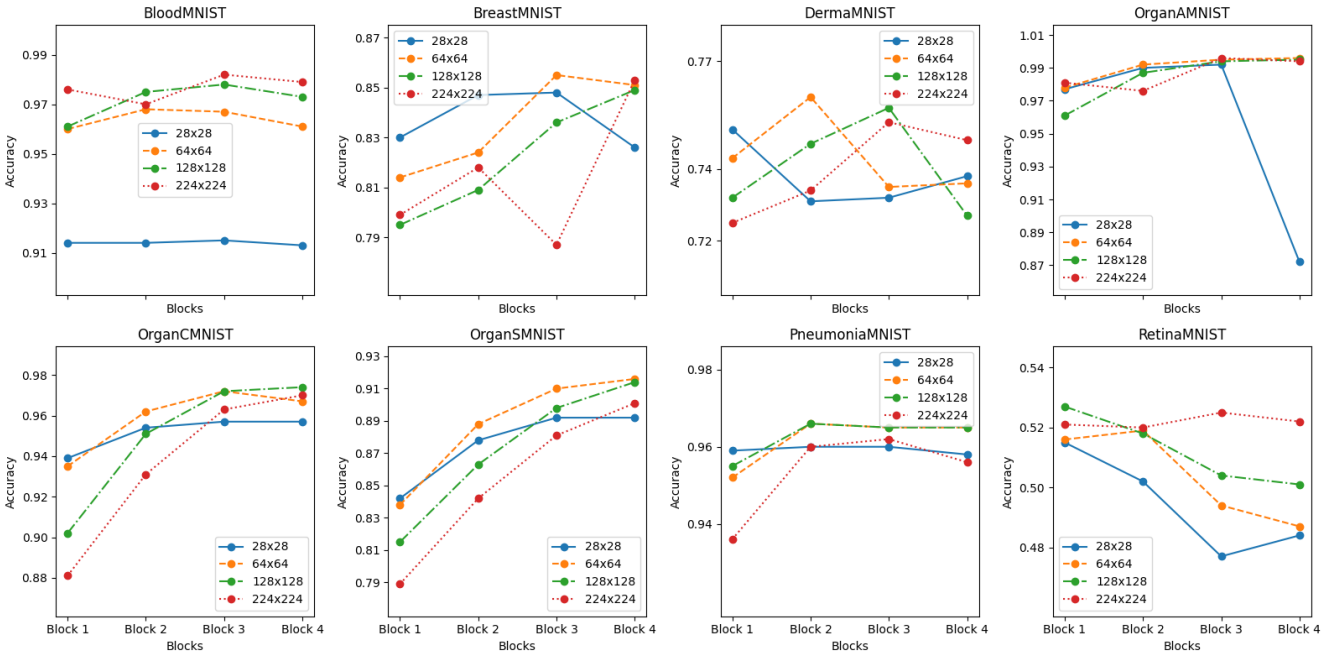

Figure 3: Evaluation of depth and resolution combinations for all datasets. The y-axis range is adjusted to highlight differences in accuracy.

one being 2x2 and the other 14x14). This suggests that there are multiple ways to extract salient features from the original images through different mathematical approaches.

Our main contribution is a systematic evaluation of network depth and image resolution combinations in medical images, a factor overlooked in related studies, including the recent one by Zhang et al. [13], which did not explore the impact of image resolution. While direct comparison is limited by different data partitions, our best models outperform theirs in accuracy for DermMNIST, OrganSMNIST, PneumoniaMNIST, and several other datasets.

Table 6: The conditions for outperforming 224x224 resolution for the block 4. The significant results are highlighted in bold.

Dataset	Resolusion	Block	FMSize	Significance
BloodMNIST	224x224	3	14x14	NS
BreastMNIST	64x64	3	4x4	NS
DermaMNIST	28x28	1	7x7	NS
DermaMNIST	**64x64**	**2**	**8x8**	p<0.05 *
DermaMNIST	128x128	3	8x8	NS
DermaMNIST	224x224	3	14x14	NS
OrganAMNIST	64x64	3	4x4	NS
OrganAMNIST	**64x64**	**4**	**2x2**	p<0.05 *
OrganAMNIST	128x128	4	4x4	NS
OrganAMNIST	**224x224**	**3**	**14x14**	p<0.05 *
OrganCMNIST	64x64	3	4x4	NS
OrganCMNIST	128x128	3	8x8	NS
OrganCMNIST	128x128	4	4x4	NS
OrganSMNIST	**64x64**	**3**	**4x4**	p<0.05 *
OrganSMNIST	**64x64**	**4**	**2x2**	p<0.001 ***
OrganSMNIST	**128x128**	**4**	**4x4**	p<0.001***
PneumoniaMNIST	28x28	1	7x7	NS
PneumoniaMNIST	28x28	2	4x4	NS
PneumoniaMNIST	28x28	3	2x2	NS
PneumoniaMNIST	28x28	4	1x1	NS
PneumoniaMNIST	**64x64**	**2**	**8x8**	p<0.01 **
PneumoniaMNIST	**64x64**	**3**	**4x4**	p<0.01 **
PneumoniaMNIST	**64x64**	**4**	**2x2**	p<0.01 **
PneumoniaMNIST	**128x128**	**2**	**16x16**	p<0.05 *
PneumoniaMNIST	**128x128**	**3**	**8x8**	p<0.01 **
PneumoniaMNIST	**128x128**	**4**	**4x4**	p<0.05 *
PneumoniaMNIST	224x224	2	28x28	NS
PneumoniaMNIST	224x224	3	14x14	NS
RetinaMNIST	128x128	1	32x32	NS
RetinaMNIST	224x224	3	14x14	NS

Table 7: Results of the average of 10 runs for different datasets (testing sets) with differing depths and resolutions. The best result for each is highlighted in bold.

Dataset - Resolution	Block1	Block2	Block3	Block4
BloodMNIST - 28x28	0.914	0.914	0.915	0.913
BloodMNIST - 64x64	0.960	0.968	0.967	0.961
BloodMNIST - 128x128	0.961	0.975	0.978	0.973
BloodMNIST - 224x224	0.976	0.970	**0.982**	0.979
BreastMNIST - 28x28	0.830	0.847	0.848	0.826
BreastMNIST - 64x64	0.814	0.824	**0.855**	0.851
BreastMNIST - 128x128	0.795	0.809	0.836	0.849
BreastMNIST - 224x224	0.799	0.818	0.787	0.853
DermaMNIST - 28x28	0.751	0.731	0.732	0.738
DermaMNIST - 64x64	0.743	**0.760**	0.735	0.736
DermaMNIST - 128x128	0.732	0.747	0.757	0.727
DermaMNIST - 224x224	0.725	0.734	0.753	0.748
OrganAMNIST - 28x28	0.977	0.990	0.992	0.872
OrganAMNIST - 64x64	0.978	0.992	0.995	**0.996**
OrganAMNIST - 128x128	0.961	0.987	0.994	0.995
OrganAMNIST - 224x224	0.981	0.976	**0.996**	0.994
OrganCMNIST - 28x28	0.939	0.954	0.957	0.957
OrganCMNIST - 64x64	0.935	0.962	0.972	0.967
OrganCMNIST - 128x128	0.902	0.951	0.972	**0.974**
OrganCMNIST - 224x224	0.881	0.931	0.963	0.970
OrganSMNIST - 28x28	0.842	0.878	0.892	0.892
OrganSMNIST - 64x64	0.838	0.888	0.910	**0.916**
OrganSMNIST - 128x128	0.815	0.863	0.898	0.914
OrganSMNIST - 224x224	0.789	0.842	0.881	0.901
PneumoniaMNIST - 28x28	0.959	0.960	0.960	0.958
PneumoniaMNIST - 64x64	0.952	0.966	0.965	0.965
PneumoniaMNIST - 128x128	0.955	**0.966**	0.965	0.965
PneumoniaMNIST - 224x224	0.936	0.960	0.962	0.956
RetinaMNIST - 28x28	0.515	0.502	0.477	0.484
RetinaMNIST - 64x64	0.516	0.519	0.494	0.487
RetinaMNIST - 128x128	**0.527**	0.518	0.504	0.501
RetinaMNIST - 224x224	0.521	0.520	0.525	0.522

Acknowledgments

The work was funded by a UWE Vice-Chancellor's Early Career Researcher Development Award. During this study, TB and HL were visiting researchers to N/Lab[6]. We are grateful to members of both N/Lab and the UWE School of Computing and Creative Technologies for many invaluable discussions. Thanks for Dr. Sue Scarborough for the advice on visualisation.

References

[1] Giansalvo Cirrincione, Rahul Ranjeev Kumar, Ali Mohammadi, Shahin Hedayati Kia, Pietro Barbiero, and Jacopo Ferretti. 2020. Shallow versus deep neural networks in gear fault diagnosis. *IEEE Transactions on Energy Conversion* 35, 3 (2020), 1338–1347.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[3] Tomoya Furukawa and Qianfgu Zhao. 2017. On Extraction of Rules from Deep Learner: The Deeper, the Better?. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*. IEEE, 54–60.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[6] Eran Malach and Shai Shalev-Shwartz. 2019. Is deeper better only when shallow is good? *Advances in Neural Information Processing Systems* 32 (2019).

[7] Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. 2016. Learning functions: when is deep better than shallow. *arXiv preprint arXiv:1603.00988* (2016).

[8] ARTZAI Picon Ruiz, AITOR Alvarez Gila, Unai Irusta, JONE Echazarra Huguet, et al. 2020. Why deep learning performs better than classical machine learning? *Dyna Ingenieria E Industria* (2020).

[9] Salvador Robles Herrera, Martine Ceberio, and Vladik Kreinovich. 2022. When is deep learning better and when is shallow learning better: Qualitative analysis. *International Journal of Parallel, Emergent and Distributed Systems* 37, 5 (2022), 589–595.

[10] Abderrezzaq Sendjasni, David Traparic, and Mohamed-Chaker Larabi. 2022. Investigating Normalization Methods for CNN-Based Image Quality Assessment. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 4113–4117.

[11] Ohad Shamir. 2018. Are resnets provably better than linear predictors? *Advances in neural information processing systems* 31 (2018).

[6] https://www.nlab.org.uk/

[12] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data* 10, 1 (2023), 41.

[13] Yu Zhang, Pengxing Cai, Yanan Sun, Zhiming Zhang, Zhenyu Lei, and Shangce Gao. 2024. A Lightweight Multi-Dendritic Pyramidal Neuron Model with Neural Plasticity on Image Recognition. *IEEE Transactions on Artificial Intelligence* (2024).