Medicine Advances

MEDICINE ADVANCES

COMMENTARY OPEN ACCESS

Overcoming Computational Resource Limitations in Deep Learning for Healthcare: Strategies Targeting Data, Model, and Computing

Han Yuan

Duke-NUS Medical School, National University of Singapore, Singapore

Correspondence: Han Yuan (yuan.han@u.duke.nus.edu)

Received: 17 October 2024 | **Revised:** 17 November 2024 | **Accepted:** 12 December 2024

Funding: The author received no specific funding for this work.

Keywords: computational resource | deep learning for healthcare | low-precision computing | model compression | sample informativeness

1 | Computational Resource Limitations

Deep learning (DL) has been identified as an indispensable backbone in health data science [1], driven by exponential growth in the scale of medical data and its remarkable modeling capability powered by increasingly complex architectures with parameter counts now surpassing hundreds of billions [2]. Supercomputing centers have been established in industry settings, but restrictions on sharing private patient data with external entities persist [3]. As a result, academic institutions and hospitals remain primary venues for developing DL models in healthcare. However, the computational constraints faced by many healthcare providers, who may lack access to high-performance computing resources, must be considered. From the perspective of the DL lifecycle, we identify three key factors that contribute to computational resource demands: data, model, and computing. Accordingly, we illustrate three representative strategies in Figure 1: informative data subset selection, model compression, and low-precision computation that offer actionable solutions for healthcare providers to mitigate computational constraints when leveraging DL models in resource-limited settings.

2 | Data Strategies

Data is the driving force behind the success of DL models. However, the rapid increase in both the dimensionality of individual samples and the overall size of datasets has made model development on entire datasets increasingly cost-prohibitive for hospitals and medical research institutes. A practical solution to this challenge is informative data subset selection, which aims to recognize pragmatic samples that would contribute to model training, thereby reducing computational costs and mitigating predictive errors caused by noisy or irrelevant data [4].

Sample selection can occur either before or during model training. Prior to training, a common approach involves representativeness-based selection, where sample representations are first obtained, followed by clustering to identify a core set of data points closest to the cluster centers [5]. During training, Katharopoulos and Fleuret [6] introduced a theory that suggests that many samples become redundant after a few epochs and can be excluded from subsequent training. Sample informativeness is measured by its contribution to the variance reduction in model parameter updates, with large batches of data being replaced by smaller subsets that maximize variance reduction. Compared with the default solution, their proposed strategy reduced the test error by 8.0% and 5.0% on CIFAR-10 and CIFAR-100 image classification tasks, respectively. Coleman et al. [7] expanded on these ideas by introducing a proxy model with fewer parameters, derived from the original backbone model, to further accelerate the calculation of sample informativeness while optimizing computational efficiency. The proposed method successfully removed 20.0% of the dataset,

Abbreviation: DL, deep learning.

https://doi.org/10.1002/med4.70001

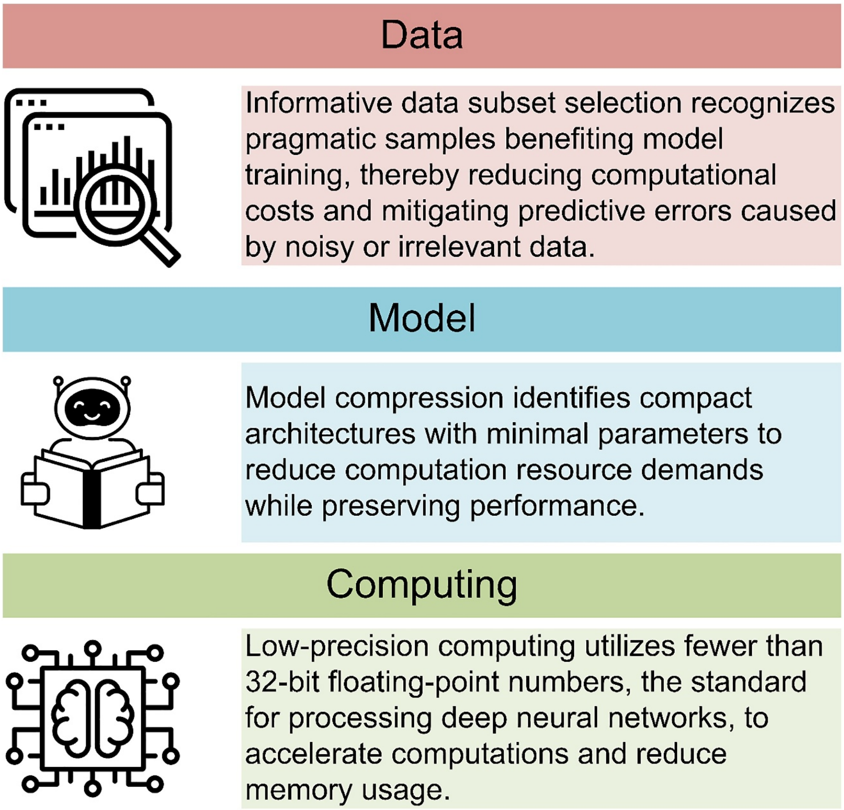

FIGURE 1 | Representative techniques toward efficient deep learning for healthcare.

with only a 0.1% increase in the top 1 error on the Amazon review polarity benchmark.

Moreover, unlike traditional DL approaches that require large datasets for training, recent advancements in foundation models across various medical domains, such as radiology [8], pathology [9], and ophthalmology [10], have demonstrated the capability to address tasks effectively with limited fine-tuning samples. This efficiency stems from their ability to leverage generalizable knowledge acquired from extensive external data on related tasks and further mitigates the potential model convergence issues caused by data subset selection [11].

3 | Model Strategies

The victory of DL originates not only from large-scale datasets but also from model parameter scalability and the corresponding capability to fit complex decision functions [12]. In the intervening time, optimizing these large-scale parameters imposes significant requirements on computational resources. Model compression aims to identify compact architectures with minimal parameters to reduce computational demands while preserving performance. Among common techniques, parameter pruning and low-rank factorization have been suggested because they can be easily deployed in both the training and inference stages [13]. Additionally, they are more efficient than techniques that require additional training, particularly in resource-constrained settings.

Parameter pruning eliminates parameters that are not important to the target task, whereas low-rank factorization leverages matrix decomposition to approximate indispensable parameters [14, 15]. For parameter pruning, Srinivas and Babu [16] proposed a seminal approach for neuron elimination. Their method calculates the saliency of each neuron and iteratively removes those with the lowest saliency. On the MNIST dataset, a LeNet-like architecture with 83.5% parameter compression achieved an accuracy of 98.4%, thereby reflecting a decrease of < 1.0% compared with the full-parameter model, which attained 99.1% accuracy. This approach can be applied to any pre-trained model without additional training, aligning with our goal of reducing computational resource consumption.

By contrast, low-rank factorization targets the matrix multiplications that are inherent in DL, which account for a substantial portion of the computational load [17]. This approach replaces the original matrix operations with decomposed matrix multiplications, thereby reducing calculation redundancy. For instance, Denton et al. [18] used singular value decomposition to approximate convolutional filter matrices with low-rank matrices, significantly reducing the number of parameters. Validated on ImageNet classification, their method achieved a reduction in weights ranging from 2.4 to 13.4 times, with a max error increase of 0.9%. However, for applications with sufficient computational resources, where the primary goal is to reduce inference latency, alternative methods such as knowledge distillation should also be considered [19]. For readers interested in a detailed exploration of model compression techniques, we recommend referring to this survey [20].

4 | Computing Strategies

In addition to data volumes and model parameters, the resource demands of hardware, storage infrastructure, and data transmission networks are significantly influenced by the computing approach adopted throughout the DL lifecycle [21]. Low-precision computing is a technique that uses fewer than 32-bit floating-point numbers, the standard for processing deep neural networks, to represent individual parameters, activations, and gradients [22]. This approach accelerates computations and reduces memory usage, thereby lowering the overall computational resource demands [23].

Low-precision computation is particularly useful during model inference [24]. Considering high-dimensional medical image data as an example, multiple images can be spliced into low-precision batches for processing by a low-precision model [24]. Then batch prediction results are separated to retrieve the output for each individual image [24]. Compared with the inference stage, using low-precision computation during training often leads to convergence-relevant issues, such as gradient divergence, vanishing gradients, or entrapment in local minima [25]. Mixed-precision computation has gained prominence as an effective strategy to address these challenges by combining low precision for less sensitive operations with sufficient numerical resolution for critical computations. Hayford et al. [25] demonstrated a practical application of mixed-precision training for DL models. In their experiments, model parameters were stored in full-precision, whereas loss and gradient parameters were computed and stored in low-precision. When model parameters were updated, these low-precision values were temporarily converted back to full-precision. Furthermore, loss scaling was used to shift the update parameters from a wide range in low-precision to a narrower range in full-precision because most activated values during training were typically < 1. The experimental results showed that mixed-precision training not only led to a relative error increase of $< 1.0\%$ compared with full-precision training but also significantly reduced computational resource consumption, with a reduction of training time ranging from 10.6% to 43.1%.

The trade-off between achieving model performance and compression cannot be fully addressed; however, various strategies have been proposed to mitigate this issue. For example, progressively decreasing the bitwidth achieves comparable or superior performance while substantially reducing the memory requirements for model parameters, compared with uniform bitwidth reduction [26]. For readers seeking an in-depth understanding of mixed-precision training and inference techniques, we recommend consulting this review [27].

5 | Toward Efficient Deep Learning

In healthcare scenarios, researchers should strike a balance between computational efficiency and the rigorous demand for accuracy and reliability: Overlooking computational efficiency can lead to impractical systems that are resource-intensive, whereas sacrificing accuracy compromises patient safety and clinical utility. This commentary elucidates the three primary factors of data, models, and computing that influence computational requirements when using DL in healthcare and presents representative solutions to reduce computational resource burdens.

Regarding future efficient DL, we recommend that researchers conduct rigorous evaluations of models, particularly compressed or low-precision models, in both out-of-sample and out-of-time scenarios. Such assessments should use diverse metrics tailored to specific objectives, including sensitivity and specificity, to accurately identify patients with and without a given disease [28, 29]. For high-level screening, such as infectious disease detection in a large population, models with high sensitivity are prioritized to minimize false negatives and mitigate the risk of disease spread [30]. By contrast, for critical diagnoses, such as confirming cancer prior to radiotherapy, models with high specificity are crucial to prevent unnecessary interventions, reduce patient anxiety, and avoid substantial costs [31]. If the performance of efficient DL solutions is validated to meet the real-world deployment standard, researchers are then encouraged to focus on computational feasibility, energy efficiency, and environmental sustainability [32, 33].

Author Contributions

Han Yuan: conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing–original draft, writing–review and editing.

Acknowledgments

The author has nothing to report.

Ethics Statement

This study is exempted from review by the ethics committee as it does not involve human participants, animal subjects, or the collection of sensitive data.

Consent

The author has nothing to report.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

The author has nothing to report.

References

1. H. Yuan, K. Yu, F. Xie, M. Liu, and S. Sun, “Automated Machine Learning With Interpretation: A Systematic Review of Methodologies and Applications in Healthcare,” *Medicine Advances* 2, no. 3 (2024): 205–237, https://doi.org/10.1002/med4.75.

2. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, “Large Language Models in Medicine,” *Nature Medicine* 29, no. 8 (2023): 1930–1940, https://doi.org/10.1038/s41591-023-02448-8.

3. H. Yuan, L. Kang, Y. Li, and Z. Fan, “Human-in-the-Loop Machine Learning for Healthcare: Current Progress and Future Opportunities in

Electronic Health Records," *Medicine Advances* 2, no. 3 (2024): 318–322, https://doi.org/10.1002/med4.70.

4. X. Xu, T. Liang, J. Zhu, D. Zheng, and T. Sun, "Review of Classical Dimensionality Reduction and Sample Selection Methods for Large-Scale Data Processing," *Neurocomputing* 328 (2019): 5–15, https://doi.org/10.1016/j.neucom.2018.02.100.

5. X. Xia, J. Liu, J. Yu, X. Shen, B. Han, and T. Liu, eds., "Moderate Coreset: A Universal Method of Data Selection for Real-World Data-Efficient Deep Learning," in *Proceedings of the International Conference on Learning Representations* (2022).

6. A. Katharopoulos and F. Fleuret, "Not All Samples Are Created Equal: Deep Learning With Importance Sampling," *in Proceedings of the 35th International Conference on Machine Learning* (2018): 1803.00942.

7. C. Coleman, C. Yeh, S. Mussmann, et al., "Selection via Proxy: Efficient Data Selection for Deep Learning," in *Proceedings of the International Conference on Learning Representations* (2020), 1–25.

8. W. F. Wiggins and A. S. Tejani, "On the Opportunities and Risks of Foundation Models for Natural Language Processing in Radiology," *Radiology Artificial Intelligence* 4, no. 4 (2022): e220119, https://doi.org/10.1148/ryai.220119.

9. R. J. Chen, T. Ding, M. Y. Lu, et al., "Towards a General-Purpose Foundation Model for Computational Pathology," *Nature Medicine* 30, no. 3 (2024): 850–862, https://doi.org/10.1038/s41591-024-02857-3.

10. G. M. Somfai, J. R. T. Zoellin, C. Merk, et al., "Evaluating the Practicability of Natural-Domain and Domain Specific Foundation Models for Ophthalmic Image Classification," *Investigative Ophthalmology & Visual Science* 65, no. 9 (2024): PB0022.

11. M. Moor, O. Banerjee, Z. S. H. Abad, et al., "Foundation Models for Generalist Medical Artificial Intelligence," *Nature* 616, no. 7956 (2023): 259–265, https://doi.org/10.1038/s41586-023-05881-4.

12. H. Yuan, "Natural Language Processing for Chest X-Ray Reports in the Transformer Era: BERT-Like Encoders for Comprehension and GPT-Like Decoders for Generation," *iRADIOLOGY* 3, no. 1 (2025): 1–8, https://doi.org/10.1002/ird3.115.

13. Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model Compression and Acceleration for Deep Neural Networks: The Principles, Progress, and Challenges," *IEEE Signal Processing Magazine* 35, no. 1 (2018): 126–136, https://doi.org/10.1109/MSP.2017.2765695.

14. P. Wimmer, J. Mehnert, and A. P. Condurache, "Dimensionality Reduced Training by Pruning and Freezing Parts of a Deep Neural Network: A Survey," *Artificial Intelligence Review* 56, no. 12 (2023): 14257–14295, https://doi.org/10.1007/s10462-023-10489-1.

15. T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and Quantization for Deep Neural Network Acceleration: A Survey," *Neurocomputing* 461, no. C (2021): 370–403, https://doi.org/10.1016/j.neucom.2021.07.045.

16. S. Srinivas and R. V. Babu, "Data-Free Parameter Pruning for Deep Neural Networks," in *Proceedings of the British Machine Vision Conference 2015* (Swansea, 2015), https://doi.org/10.5244/c.29.31.

17. Z. Hu, F. Nie, R. Wang, and X. Li, "Low Rank Regularization: A Review," *Neural Networks* 136 (2021): 218–232, https://doi.org/10.1016/j.neunet.2020.09.021.

18. E. Denton, W. Zaremba, J. Bruna, et al., "Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation," in *Proceedings of the 28th International Conference on Neural Information Processing Systems – Volume 1* (Montreal, Canada, 2014), 1269–1277, https://doi.org/10.5555/2968826.2968968.

19. L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey," *Proceedings of the IEEE* 108, no. 4 (2020): 485–532, https://doi.org/10.1109/JPROC.2020.2976475.

20. T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, "A Comprehensive Survey on Model Compression and Acceleration," *Artificial Intelligence Review* 53, no. 7 (2020): 5113–5155, https://doi.org/10.1007/s10462-020-09816-7.

21. Z. Jia, J. Chen, X. Xu, et al., "The Importance of Resource Awareness in Artificial Intelligence for Healthcare," *Nature Machine Intelligence* 5, no. 7 (2023): 687–698, https://doi.org/10.1038/s42256-023-00670-0.

22. J. Bjorck, X. Chen, C. De Sa, et al., "Low-Precision Reinforcement Learning: Running Soft Actor-Critic in Half Precision," *in Proceedings of the 38th International Conference on Machine Learning* (2021): 2102.13565.

23. Z. Cai, X. He, J. Sun, and N. Vasconcelos, "Deep Learning With Low Precision by Half-Wave Gaussian Quantization," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI, 2017), 5406–5414, https://doi.org/10.1109/CVPR.2017.574.

24. C. Yu, Y. Liu, X. Xia, D. Lan, X. Liu, and S. Wu, "Precise and Fast Segmentation of Offshore Farms in High-Resolution SAR Images Based on Model *Fusi* on and Half-Precision Parallel Inference," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022): 4861–4872, https://doi.org/10.1109/JSTARS.2022.3181355.

25. J. Hayford, J. Goldman-Wetzler, E. Wang, and L. Lu, "Speeding up and Reducing Memory Usage for Scientific Machine Learning *via* Mixed Precision," *Computer Methods in Applied Mechanics and Engineering* 428 (2024): 117093, https://doi.org/10.1016/j.cma.2024.117093.

26. T. Chu, Q. Luo, J. Yang, and X. Huang, "Mixed-Precision Quantized Neural Networks With Progressively Decreasing Bitwidth," *Pattern Recognition* 111 (2021): 107647, https://doi.org/10.1016/j.patcog.2020.107647.

27. M. Rakka, M. E. Fouda, P. Khargonekar, and F. Kurdahi, "A Review of State-of-the-Art Mixed-Precision Neural Network Frameworks," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, no. 12 (2024): 7793–7812, https://doi.org/10.1109/TPAMI.2024.3394390.

28. H. Yuan, "Toward Real-World Deployment of Machine Learning for Health Care: External Validation, Continual Monitoring, and Randomized Clinical Trials," *Health Care Science* 3, no. 5 (2024): 360–364, https://doi.org/10.1002/hcs2.114.

29. A. Swift, R. Heale, and A. Twycross, "What Are Sensitivity and Specificity?," *Evidence-Based Nursing* 23, no. 1 (2020): 2–4, https://doi.org/10.1136/ebnurs-2019-103225.

30. H. R. Chiu, C.-K. Hwang, S.-Y. Chen, et al., "Machine Learning for Emerging Infectious Disease Field Responses," *Scientific Reports* 12, no. 1 (2022): 328, https://doi.org/10.1038/s41598-021-03687-w.

31. N. Banaei, J. Moshfegh, A. Mohseni-Kabir, J. M. Houghton, Y. Sun, and B. Kim, "Machine Learning Algorithms Enhance the Specificity of Cancer Biomarker Detection Using SERS-Based Immunoassays in Microfluidic Chips," *RSC Advances* 9, no. 4 (2019): 1859–1868, https://doi.org/10.1039/c8ra08930b.

32. L. Bouza, A. Bugeau, and L. Lannelongue, "How to Estimate Carbon Footprint When Training Deep Learning Models? A Guide and Review," *Environmental Research Communications* 5, no. 11 (2023): 115014, https://doi.org/10.1088/2515-7620/acf81b.

33. H. Yuan, "Agentic Large Language Models for Healthcare: Current Progress and Future Opportunities," *Medicine Advances* 3, no. 1 (2025), https://doi.org/10.1002/med4.70000.