**Initial Post: Self-supervised learning – general concepts and applications within medicine**

Machine learning (ML) is the process by which computer algorithms are built to perform tasks without being specifically told how to, and to improve with experience. ML is generally divided into supervised learning, where models learn to predict an outcome based on input data and a human-generated output label, and unsupervised learning, where the algorithm tries to identify relevant underlying patterns in the input data without any labels (Russel and Norvig, 2021).

Alternative ML paradigms are also emerging, including self-supervised learning (SSL). In SSL, the model learns relevant features about input, unlabelled data by generating informative labels itself using “pretext” tasks, with these for supervised learning tasks. The model is then retrained or refined in small sets of labelled data (Krishnan, Rajpurkar and Topol, 2022).

The initial step can be performed using two major approaches (Huang, Voorham and Haaijer-Ruskamp, 2016). In contrastive learning, the model learns to associate related inputs and contrast unrelated inputs. Training data can include groupings of naturally associated inputs (e.g. different pictures of one individual, or perspectives from the same medical imaging study), or artificial variations generated by data augmentation (e.g. cropping, rotating, or adding noise to images) (Yuan et al., 2024). In generative learning, the model is trained to predict a particular section of an input (such as an image or text chunk) which has been blanked out by learning about adjacent or contextual inputs (Ohri and Kumar, 2021).

SSL is particularly valuable for tasks where labelling is costly or time-consuming, as is frequent in medicine, such as interpreting histology slides, CT scan images, or cardiac monitoring data (Krishnan, Rajpurkar and Topol, 2022). Applications of SSL include parsing EHR to predict disease, and identifying pathologies from fundal images, X-ray images, or histology slides (Krishnan, Rajpurkar and Topol, 2022). Importantly, SSL has been shown to meet or exceed performance of purely supervised learning models, possibly since learning is based only on informative underlying features rather than spurious associations (e.g. using backgrounds instead of actual subjects in image classification, i.e. shortcut learning) (Geirhos et al., 2020; Ohri and Kumar, 2021). SSL may also be better suited for multi-modal data by using associative learning (e.g. matching X-ray images with corresponding text reports) (Krishnan, Rajpurkar and Topol, 2022).

However, this ML paradigm has some important weaknesses, namely in relation to identifying useful pretext tasks, the impact of choice of data augmentation procedures on model performance, and its relative novelty and lack of implementation guidelines, especially with respect to medical data (Ohri and Kumar, 2021; Krishnan, Rajpurkar and Topol, 2022).

References:

Geirhos, R. et al. (2020) ‘Shortcut learning in deep neural networks’, *Nature Machine Intelligence*, 2(11), pp. 665–673. Available from: https://doi.org/10.1038/s42256-020-00257-z.

Huang, Y., Voorham, J. and Haaijer-Ruskamp, F.M. (2016) ‘Using primary care electronic health record data for comparative effectiveness research: experience of data quality assessment and preprocessing in The Netherlands’, *Journal of Comparative Effectiveness Research*, 5(4), pp. 345–354. Available from: https://doi.org/10.2217/cer-2015-0022.

Krishnan, R., Rajpurkar, P. and Topol, E.J. (2022) ‘Self-supervised learning in medicine and healthcare’, *Nature Biomedical Engineering*, 6(12), pp. 1346–1352. Available from: https://doi.org/10.1038/s41551-022-00914-1.

Ohri, K. and Kumar, M. (2021) ‘Review on self-supervised image recognition using deep neural networks’, *Knowledge-Based Systems*, 224, p. 107090. Available from: https://doi.org/10.1016/j.knosys.2021.107090.

Russel, S. and Norvig, P. (2021) *Artificial intelligence: A modern approach*. Global Edition. Pearson Higher Education.

Yuan, H. et al. (2024) ‘Self-supervised learning for human activity recognition using 700,000 person-days of wearable data’, *npj Digital Medicine*, 7(1), p. 91. Available from: https://doi.org/10.1038/s41746-024-01062-3.