**Peer response: Supervised and Unsupervised ML Algorithms in Disease Detection**

Masana discusses machine learning (ML) in disease detection. She addresses supervised learning with K-nearest neighbours (KNN), which identify disease based on similarities with labelled examples. Proposed advantages are ease of implementation and interpretability, and dependence on labelled data the major disadvantage.

Mateusz argues that KNN models could become problematic given their over-reliance on training data, potentially cause oversimplification of complex phenomena by namely mistaking drastically different diseases with similar clinical presentations (e.g. cough in common cold vs tuberculosis). This is, in my view, incorrect. First, the major limitation of supervised learning is (un)availability of labelled data, not labels themselves (although incorrect labels would obviously compromise performance) (Yakimovich, Beaugnon, Huang and Ozkirimli, 2021). Second, symptom checkers generally consider other features besides isolated symptoms, namely duration, presence of accompanying features (blood on cough), and other symptoms (fever, weight loss).

Martyna supported the value of KNN, adding that alternative algorithms such as decision trees yield poor results for “minority groups” (i.e. unbalanced data classes). However, the same may be true of KNN, as the algorithm may wrongly tend to pick the majority class (Shi, 2020).

On unsupervised learning, Masana writes that K-means clustering (KMC) detect unlabelled outliers within a dataset, helping identify “unfamiliar diseases”. She correctly identifies challenges with k choice and assumption of spherical clustering. Mateusz suggests KMC may outperform KNN as they prevent oversimplification, and better handle “multifactor symptoms” (the meaning of which is unclear; many symptoms? non-specific? symptoms with complex features?). He also suggested using (one-class) support vector machines, which identify whether a new instance is an outlier versus those in the training data (Dreiseitl, Osl, Scheibböck and Binder, 2010). While unsupervised learning approaches can indeed be helpful at identifying outliers (“unknown diseases”), I argue they are better used for unravelling unknown relationships within a particular condition, for example by clustering patients according to presentation, risk factors, or disease progression (Segar et al., 2020).

Paul proposed two models, Faster R-CNN and VGG (Visual Geometry Group), supposed to blend supervised and unsupervised learning. However, according to his reference, these models were developed for image classification, and it is unclear why they would outperform those previously mentioned (or what “splitting the algorithms to hybrids” means). Given their higher complexity and unclear advantages, I argue against their application here.

Masana ends with “choice of algorithm [supervised vs unsupervised] depends on data available”. I would counter-argue that choice of ML paradigm/model should first consider the problem to be addressed. Data should then be collected (and engineered) based on this, with model selection/tuning as the last (crucial) step (Duboue, 2020).

References:

Dreiseitl, S., Osl, M., Scheibböck, C. and Binder, M. (2010) ‘Outlier Detection with One-Class SVMs: An Application to Melanoma Prognosis’, *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2010, pp. 172–176.

Duboue, P. (2020) ‘The Art of Feature Engineering’, in. Cambridge University Press.

Segar, M.W. et al. (2020) ‘Phenomapping of patients with heart failure with preserved ejection fraction using machine learning-based unsupervised cluster analysis’, *European Journal of Heart Failure*, 22(1), pp. 148–158. Available from: https://doi.org/10.1002/ejhf.1621.

Shi, Z. (2020) ‘Improving k-Nearest Neighbors Algorithm for Imbalanced Data Classification’, *IOP Conference Series: Materials Science and Engineering*, 719(1), p. 012072. Available from: https://doi.org/10.1088/1757-899X/719/1/012072.

Yakimovich, A., Beaugnon, A., Huang, Y. and Ozkirimli, E. (2021) ‘Labels in a haystack: Approaches beyond supervised learning in biomedical applications’, *Patterns*, 2(12), p. 100383. Available from: https://doi.org/10.1016/j.patter.2021.100383.

Decision-trees may be better suited for disease classification (easier explainability)

inefficient and costly, especially for large and high-dimensional datasets

perform poorly if data are categorical

KNN does not have a training step; data need to be read for each object that is added; slow with large datasets, curse of dimensionality, sensitivity to outliers, class imbalance