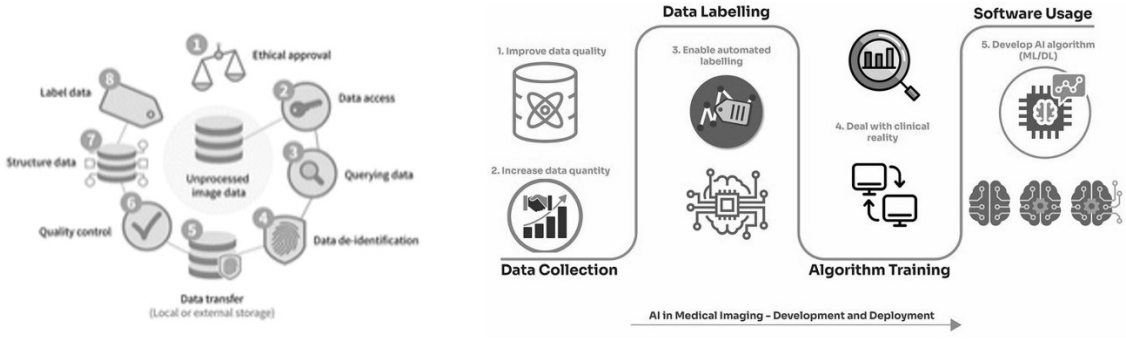
Improve Medical Diagnosis with Data Science

Medical diagnosis is the process of determining a person's health condition based on the symptoms and signs as a basis for taking specific medical care. Therefore, medical diagnosis is an essential aspect of medical care. The accuracy of the diagnosis at the right time is critical to prevent inappropriate medical care. Errors in diagnosis (misdiagnosis) can occur when medical professionals make decisions and actions that do not solve the patient's problem (cure patient disease). A late diagnosis can also occur when medical professionals are late to identify a patient's condition, which causes delays in medical care. These two conditions can occur even for competent and experienced medical professionals.

Approximately 5% of adult patients in the United States are misdiagnosed annually, with over 12 million people. Research shows that about 10% of cases of patient death in the United States are due to patients' misdiagnosis (National Academies of Science Engineering and Medicine, 2015). Errors or delays in diagnosis can occur because of the abundance of medical data that needs to be *patterned* and interpreted according to certain medical conditions. Medical data in this discussion is the medical image of certain medical conditions (such as CT scans, MRIs, and X-rays). However, extracting qualitative information from medical imaging from a huge dataset would be time-consuming, tedious, and practically impractical. Therefore, we need a program that automatically analyses medical images into qualitative information. Hence the repetitive work to determine specific *patterns* in medical images and interpret those into qualitative information will be done by a program. That is why data science can join *the game*.

Deep learning is a part of data science that can be used to develop program models to interpret images into qualitative information. In developing the model, we can follow the flowchart diagrams in Figure 1. Based on Figure 1, we can conclude that the process requires building a data science model in medical imaging. The first step is to obtain ethical approval from the authority. After that, it is necessary to access and gather medical image datasets. The next step is to improve data quality and increase the quantity. Then, label the data and relate it to specific diseases. After that, we can develop deep learning algorithms to recognise medical image conditions to their disease and symptoms. The final step is to evaluate the process to

determine ambiguities or biases from the output. After the evaluation, the next step is to deploy the program.



A B

Figure 1 (A) Process of medical image data handling (Willemink et al., 2020); (B) Pathway of artificial intelligence in medical imaging (Hameed et al., 2021)

A program model like this can make diagnoses based on specific medical images more quickly and accurately. However, this program model will be more accurate if the dataset used is enormous. Thus, all the probabilities of certain medical conditions can be known in addition to using images to aid diagnosis related to medical imagery. With scientific data, a diagnosis can also be made based on the symptoms experienced by the patient.

The process that needs to be done to create a model of a diagnosis program based on symptoms can be carried out with a procedure similar to image interpretation. The difference is the dataset used, where datasets are needed to be related to certain diseases and their symptoms. Next, it is necessary to determine the algorithmic approach to be used. There are two commonly used approaches: the standard associative algorithm (Figure 2A) and the counterfactual algorithm (Figure 2B). However, considering the accuracy of the diagnosis, the counterfactual approach has better accuracy. In contrast, the associative approach has a diagnostic accuracy in the top 48%, while the counterfactual approach is in the top 25%.

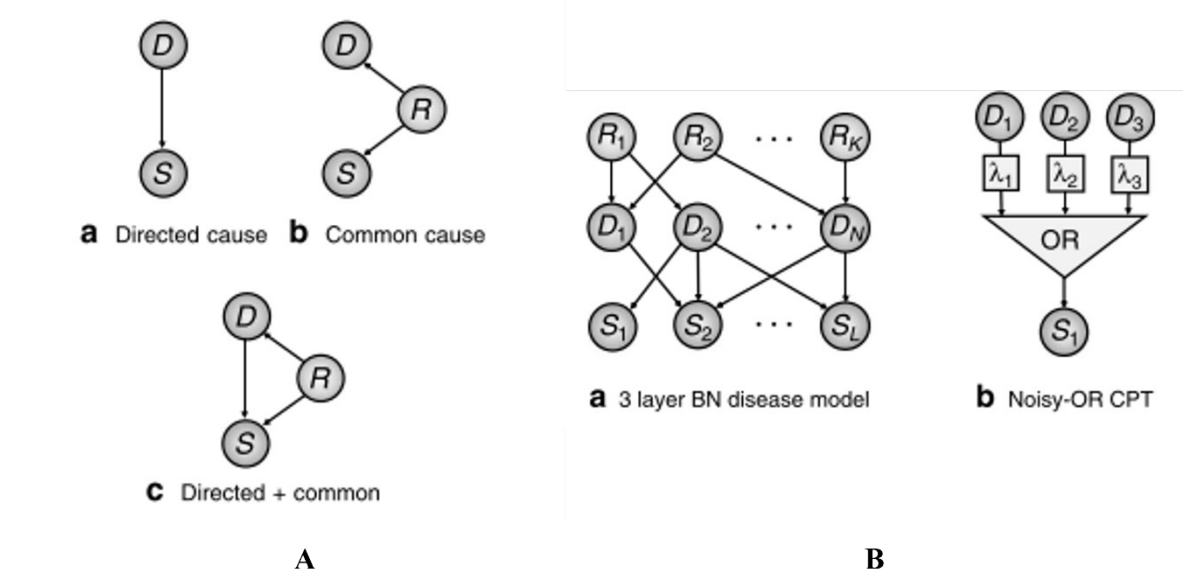


Figure 2 (A) Three distinct causal structures for diseases and symptoms. **a** Disease D is a direct cause of symptom S, **b** D does not cause symptom S, but they are correlated by a latent common cause R, **c** D is a direct cause of S, and a latent common cause R is present; (B) Generative structure of our diagnostic Bayesian networks. **a** Three-layer Bayesian network representing risk factors Ri, diseases Dj and symptoms Sk. **b** noisy-OR CPT. S is the Boolean OR function of its parents, each with an independent probability λi of being ignored, removing them from the OR function (Richens et al., n.d.)

Data science and deep learning can improve medical diagnosis` accuracy and efficiency. This model is designed to recognise health problems using medical images and symptoms that have better accuracy and efficiency than medical professionals. The impact of this model will be positive in the health care industry.

References

Hameed, B. M. Z., Prerepa, G., Patil, V., Shekhar, P., Raza, S. Z., Karimi, H., Paul, R., Naik, N., Modi, S., Vigneswaran, G., Rai, B. P., Chłosta, P., & Somani, B. K. (2021). *Engineering and clinical use of artificial intelligence (AI) with machine learning and data science advancements : radiology leading the way for future*. 1–12. https://doi.org/10.1177/17562872211044880

National Academies of Science Engginering and Medicine. (2015). *Improving Diagnosis in*

Health Care (E. P. Balogh, B. T. Miller, & J. R. Ball (eds.)). The National Academies Press. https://doi.org/10.17226/21794

Richens, J. G., Lee, C. M., & Johri, S. (n.d.). Improving the accuracy of media diagnosis with causal machine learning. *Nature Communications*, *2020*, 1–9. https://doi.org/10.1038/s41467-020-17419-7

Willemink, M. J., Koszek, W. A., Hardell, M. S. C., Wu, M. S. J., Rubin, D. L., & P, M. S. M. (2020). *Preparing Medical Imaging Data for Machine Learning*. *21*.

GitHub Repository Link