**Book**: XAI: Interpreting, Explaining, and Visualizing Deep Learning – State of the Art Survey.

Reading the scoring System – Batarseh

Read the paper on XAI Method 10 vs 9 DL

Apply SHAP on simple methods – Pytorch

-----------------------------------------------------------------------------------------------------------------------------------’

**References**

1. Alber, M.: Software and application patterns for explanation methods. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS, vol. 11700, pp. 399–433, Springer, Cham (2019)

2. Alber, M., et al.: iNNvestigate neural networks!. J. Mach. Learn. Res. 20, 1–8 (2019)

3. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Gradient-based attribution methods. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS, vol. 11700, pp. 169–191, Springer, Cham (2019)

4. Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE ICASSP, pp. 6645–6649 (2013)

5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems, pp. 1106–1114 (2012)

6. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.-R.: Layer-wise relevance propagation: An overview. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS, vol. 11700, pp. 193–209, Springer, Cham (2019)

7. Schütt, K.T., Arbabzadah, F., Chmiela, S., Müller, K.-R., Tkatchenko, A.: Quantumchemical insights from deep tensor neural networks. Nat. Commun.

8, 13890 (2017) 8. Thomas, A.W., Heekeren, H.R., Müller, K.-R., Samek, W.: Analyzing neuroimaging data through recurrent deep learning models (2018). arXiv preprint arXiv:1810.09945

-----------------------------------------------------------------------------------------------------------------------------------

**1/ Preface:**

* Before deploying an AI system, we see a strong need to validate its behavior, and thus establish guarantees that it will continue to perform as expected when deployed in a real-world environment.
* More recent deep learning based neural networks provide far superior predictive power, but at the price of behaving as a ‘black-box’ where the underlying reasoning is much more difficult to extract