**Preliminary Interpretability Questionaire for the SLENDER project**

The preliminary unified scoring system to estimate the added value of any cNN and cDEBI-NN interpretabiltiy approach implemented throughout the SLENDER project. This table may change and be further improved throughout over time.

|  |  |  |
| --- | --- | --- |
| **Question** | **Subject** | **Metrics** |
| Does the given interpretation demonstrate if the input dataset was sufficient for the training? | Network | Yes/No |
| Does the given interpretation approach demonstrate if the training set have outliers or high imbalance ratios? | Network | Yes/No |
| Does the given interpretation approach imply whether a different model configuration should have been utilized for training? | Network | Yes/No |
| How easy it is to get an overview of the model configuration (properties of convolutional and fully connected layers) | Network | 1(hard) – 5 (easy) |
| How many convolutional and hidden layers does the network have? | Network | Provided values |
| How clear is it whether the network has a balanced weight distribution and if it could be further simplified by e.g. a simpler model scheme or more sparsity? | Network | 1(very unclear) – 5 (very clear) |
| Are there corrupted neuron input, weight and/or activation function configurations in the network that would imply a suboptimal training? | Network | Yes/No |
| How easily does the given interpretation approach demonstrate if the given inference input belonged to the minority subgroup? | Inference | 1(hard) – 5 (easy) |
| Does the given interpretation approach imply prediction certainty metrics, considering the training data characteristics and the relationship of the inference data to it? | Inference | Yes/No |
| Is it clear what parts of the input data are relevant for training | Inference | 1(very unclear) – 5 (very clear) |
| Does the network provide its prediction based on clinically-relevant patterns in the inference data? | Inference | Yes/No |
| Hoes the given interpretability approach help to understand and justify why the given network configuration was resulting in its prediction performance? | Network | Yes/No |
| Goes the given interpretability approach provide any new information or value related to the properties of the training data? If yes, specify. | Network | Yes/No + detailed answer |
| Can you explain why the given prediction was provided for the given inference data in relation to the given interpretability approach? | Inference | Yes/No + detailed answer |
| How much do you trust the given model, based on the provided interpretability approach? Specify why. | Network / Inference | 1(do not trust) – 5(highly-trust) + detailed answer |
| Explain why you accept or do not accept the output of the model based on the given interpretability approach | Inference | Detailed answer |