Deep-Learning Radiomics Research Checklist (DLRRC)

| No. | Item with Answers & Scores |
| --- | --- |
| **Part I: Basic information and Data Acquisition (10 points)** | |
| 1 | Labels are meaningful and biological discrepancy, mentioning potential topological differences existing.  *Answers and scores: Yes (2.5); No (0).* |
| 2 | Filtration of radiological data are applied and described.  *Answers and scores: Yes (2.5); No (0).* |
| 3 | Radiological data types:   * If several modalities are involved, each type of modality should have an acceptable ratio with detailed descriptions. * If only one modality is involved, declaration is required.   *Answers and scores: Yes (2.5); No (0).* |
| 4 | For data sources:   * If it is all originated from open sources, situation of application and filters should be declared. * If it involves closed-source data, institutional ethical reviewing approvement and serial numbers are required.   *Answers and scores: Yes (2.5); No (0).* |
| **Part II: Pre-processing (27.5 points)** | |
| 5 | The ratio of each label should be unextreme and acceptable.  *Answers and scores: Yes (2.5); No (0).* |
| 6 | The ratio of generated images should be closed, otherwise reasonable explanation is required.  *Answers and scores: Yes (2.5); No (0).* |
| 7 | Processing staffs are radiological professionals.  *Answers and scores: Yes (2.5); No (0).* |
| 8 | Pre-processing is processed by the gold standard guidance.  *Answers and scores: Yes (2.5); No (0).* |
| 9 | Effective methods exist to promise accuracy of pre-processing.  *Answers and scores: Yes (2.5); No (0).* |
| 10 | The ratio of each dataset is reasonable.  *Answers and scores: Yes (2.5); No (0).* |
| 11 | Cases are independent, which aren’t involved in different datasets.  *Answers and scores: Yes (2.5); No (0).* |
| 12 | Datasets are established by random selection, without manual manipulating.  *Answers and scores: Yes (2.5); No (0).* |
| 13 | Examples of pre-processing are listed.  *Answers and scores: Yes (2.5); No (0).* |
| 14 | Methods of data augmentation are suited and correctly applied.  *Answers and scores: And (5); Or (2.5); Nor (0).* |
| **Part III: Model and Dependence (10 points)** | |
| 15 | Methods to avoid overfitting are applied and described.  *Answers and scores: Yes (2.5); No (0).* |
| 16 | Software environment should be listed, like serial numbers of version.  *Answers and scores: Yes (2.5); No (0).* |
| 17 | Hardware details should be listed.  *Answers and scores: Yes (2.5); No (0).* |
| 18 | Applied models are suited for tasks.  *Answers and scores: Yes (2.5); No (0).* |
| **Part IV: Training, Validation and Testing (27.5 points)** | |
| 19 | Training details are listed, like epoch and time spent.  *Answers and scores: Yes (2.5); No (0).* |
| 20 | Hyperparameters are listed (at least including batch size and learning rate).  *Answers and scores: Yes (2.5); No (0).* |
| 21 | The curves of accuracy-epoch and loss-epoch are provided.  *Answers and scores: Yes (2.5); No (0).* |
| 22 | The end of training is decided by the performance trends in validation datasets and is described.  *Answers and scores: Yes (2.5); No (0).* |
| 23 | Methods to promote robustness are applied.  *Answers and scores: Yes (2.5); No (0).* |
| 24 | The details of initial weights are described.  *Answers and scores: Yes (2.5); No (0).* |
| 25 | Training workloads are listed.  *Answers and scores: Yes (2.5); No (0).* |
| 26 | Objective indexes are reasonable and complete.  *Answers and scores: And (5); Or (2.5); Nor (0).* |
| 27 | Comparison between testing performance and manual processing with the same test dataset is provided.  *Answers and scores: Yes (2.5); No (0).* |
| 28 | Details of test datasets and speculation results are provided.  *Answers and scores: Yes (2.5); No (0).* |
| **Part V: Accessibility (25 points)** | |
| 29 | Codes.  *Answers and scores: Fully accessible (10); Partly accessible (7.5); Possible accessible (5); Not (0).* |
| 30 | Datasets.  *Answers and scores: Fully accessible (10); Partly accessible (7.5); Possible accessible (5); Not (0).* |
| 31 | Weights.  *Answers and scores: Accessible (5); Inaccessible (0).* |

*Item 1 Labels are meaningful and biological discrepancy, mentioning potential topological differences existing.* Labels are expected for meaning, so as to match the needs of clinic & radiology. Considering the black box of training, labels should be different in pathology or physiology, for a better interpretability and rationality. For example: 1) different pathological subtypes in classification tasks; 2) different tissues in segmentation tasks, like tumors and normal tissue.

*Item 2 Filtration of radiological data are applied and described.* The radiological data should be checked for quality and phases after data acquisition. Cases with poor quality or are noisy should be eliminated. Elimination of cases should be declared in reports. The adjustment of scales for adapting models and data balancing is encouraged to declare in articles.

*Item 3 Radiological data types*. We require authors to provide clear application of modals. If several modals are used in a report, authors should declare specific application circumstances of each modal. Difference between modal ratios of cases can be potential cause of incorrect feature extraction.

*Item 4 Data source.* Detailed application should be listed if raw data involves open source dataset, like TCIA. The detailed application should include but are not limited to inclusion/exclusion criteria and repertoires of applied cases. Institutional ethical reviewing and approval number are required if non-open data is used.

*Item 5 The ratio of each label should be unextreme and acceptable.* We suggest that authors should apply appropriate datasets for model establishment. The “appropriate” term is defined as close proportions of labels. Extreme and imbalance datasets will ruin model performance. For example: 1) in a dichotomy research, the ratio of type A and type B should be close if special imbalance-related strategy doesn’t exist. If several subtypes are involved in type A, ratios of these subtypes should be close too. 2)in a multi-label research, ratio of each label should be close. Definition of “close” is the ratio between any two types in datasets should be less than 200% and higher than 50%.

*Item 6 The ratio of generated images should be closed, otherwise reasonable explanation is required.* We don’t suggest methods that balance datasets by adjustment of generating images in imbalance datasets. If type A have 40 cases and type B have 20 cases, ratios of their generated images should also be close to 40:20 when case sizes are similar. However, some existed reports adjust imbalance datasets in this method: they generate less images from type A and more images from type B, causing close quantity in images. This method is controversial for possible feature losing of type A, which isn’t encouraged.

*Item 7 Processing staffs are radiological professionals.* Pre-processing is processed with radiological staffs, which can be regarded as standard operations. Authors should describe these details and list staffs in corresponding authors.

*Item 8 Pre-processing is processed by the gold standard guidance.* We suggest pre-processing is processed with existing gold standard results, which can partly promise interpretability and consistency of ground truth and training datasets. Considering for operational feasibility, it is comprehensible that process with visible features or macroanatomy but unsupported in the future, that is why we don’t score for this method.

*Item 9 Effective methods exist to promise accuracy of pre-processing.* Considering for potential effects of manual pre-processing, we require researchers to apply necessary methods to reduce unanticipated impacts, like cross-validation in pre-processing.

*Item 10 The ratio of each dataset is reasonable.* We require reasonable ratios of training/validation/testing datasets. It is unacceptable if testing datasets and validation datasets are small than 10% of all. Performance assessment will be incomplete and risky with too small testing datasets. If small testing dataset must use, necessary explanation is required.

*Item 11 Cases are independent, which aren’t involved in different datasets.* We require authors do not establish datasets with selection of generated images. Instead, datasets should be established with selection of cases. If images from one case are involved in training and testing datasets, performance indexes will be inflating and robustness will be damaged.

*Item 12 Datasets are established by random selection, without manual manipulating.* We don’t suggest dispensable manual processing in DL-Radiomics, which will damage robustness and reproducibility possibly. Manual manipulating in dataset establishment is not encouraged, as a method that possible affect reproducibility.

*Item 13 Examples of pre-processing are listed.* Pre-processing examples can be an evidence of high-quality data in reports. Also, imperfections or mistakes of these processes can be easily found in figures.

*Item 14 Methods of data augmentation are suited and correctly applied.* Data augmentation is widely used in DeepLearning. However, not every methods are suited for Radiomics and Translational Medicine. We don’t suggest applying methods that possibly cause pixel information changing and fake feature generation, like Gaussian Noise. Also, overaggressive strategy is not encouraged, as no meaningful improvement in performance.

*Item 15 Methods to avoid overfitting are applied and described.* We encourage some new methods and strategy application to avoid over-fitting, like some methods of escaping saddle points. The reason of choosing these method are encouraged to explain in article. Appropriate end of training can be admitted as an effective method to avoid over-fitting.

*Item 16 Software environment should be listed, like serial numbers of version.* We require authors to provide detailed information about dependent software, which isn’t hard for a real research. Not only Python and CUDA/cuDNN versions, details of Tensorflow/PyTorch and package are required too. It is admitted if these information are listed in open-access code package and can be searched.

*Item 17 Hardware details should be listed.* Hard-wares are important in DL model deployment. These information can be used in authenticity and reproducibility assessment with other items and manufacturer information.

*Item 18 Applied models are suited for tasks.* The “suited” term is defined as: 1) applied models are suited for research tasks. For example, ResNet is suited for two-label classification tasks. 2) Applied models are suited for data size. For example, ResNet-18 is suited for small datasets, and ResNet-101 is not suited. 3) Applied models can be deployed in hardware environment. For example, 3D U-Net usually can’t deploy in workstations with small graphic memory.

*Item 19 Training details are listed, like epoch and time spent.* We encourage authors to provide training details, including epochs of training, spent time in each epoch and a total spent time in training. These details can be used in reproduction and authenticity assessments.

*Item 20 Hyperparameters are listed (at least including batch size and learning rate).* We require authors to provide detailed hyper-parameters in articles, which are important in model training and non-technical sensitive. The baseline is that batch size and learning rate are provided in articles.

*Item 21 The curves of accuracy-epoch and loss-epoch are provided.* These curves can be evaluation basis of model well-trained, and also play an important role in reproduction and authenticity assessment.

*Item 22 The end of training is decided by the performance trends in validation datasets and is described.* As a “black box” process, model training situation can’t measure directly and is usually speculated by trends of performance variation in validation datasets or internal testing datasets. We suggest authors to decide training end with accuracy/loss-epoch curves.

*Item 23 Methods to promote robustness are applied.* Robustness promotion methods are encouraged to apply in research. For example, cross-validation in annotation and 10-fold cross-validation in training can be regarded as effective methods in robustness promotion.

*Item 24 The details of initial weights are described.* Initial weights are required to be declared in articles, which is meaningful for reproducibility. For transfer learning, we suggest authors to provide initial weights if pre-training is applied in open source data like Cifar-100 and ImageNet.

*Item 25 Training workloads are listed.* We suggest authors to provide workloads of training, which can be used to assess reproducibility with other items. Also, for some reports targeted in edge deployment, training workloads is a key index for feasiblity evaluation.

*Item 26 Objective indexes are reasonable and complete.* We require authors to provide suitable objective indexes about model performance. These indexes should be suitable for tasks. For example, F1 score is suitable in mutli-label classification tasks, but is not suitable in dichotomy classification tasks.

*Item 27 Comparison between testing performance and manual processing with the same test dataset is provided.* We encourage authors to provide comparisons between model testing performance and manual processing performance in testing dataset. It can be used to prove the normalization of testing datasets, by checking the manual performance and existed reports.

*Item 28 Details of test datasets and speculation results are provided.* We suggest authors to provide testing dataset details and result details in articles, involving the false cases and possible reasons of mistakes.

*Item 29 Code.* We require authors to provide accessibility of codes. Codes should be acquirable by websites (like github) or additional materials of reports. For some reports belonging to ongoing research, it is understandable but don’t score that codes only can be acquired by connecting with corresponding authors. Considering possible restrictions in code opening, it is acceptable if the technological details are described in articles. The score of this item is connected to the accessibility of codes. There are four situation of code accessibility: 1) Fully accessible. Codes are open-access in webistes or additional materials. 2) Partly accessible. Codes are not open. Authors describe detailed strucutres and modifications of models in articles. 3) Possible accessible. Codes are not open, and authors don’t describe detailed information about applied models. 4) Not accessible. Authors don’t mention about models and relative information in paper.

*Item 30 Datasets.* We require authors to provide some information and accessibility about datasets in reports. It should be fully accessible if it is all originated from a open source(like TCIA). If the raw data includes some cases from closed sources, the variety of accessibility is connected to the scores. There are four situation of dataset accessibility: 1)Fully accessible. Raw data and after-processing data are open-access. 2) Partly accessible. Cases from open source are described, involving specific application lists and inclusion/exclusion criteria. Data from closed source are not open. Authors provide detailed description of closed data, involving general condition of cohorts, institutional ethical reviewing and approval number. Also, after-processing examples are required. 3)Possible accessible. Comparing with Partly accessible, this situation are lack of after-processing examples and detailed information of open-source application. 4) Not accessible. Authors don’t provide enough information to access datasets.

*Item 31 Weights.* The accessibility of weights is one of the most factor in reproduction. It is a main point in reproducibility assessments.