**Review protocol**

The ICCV and MICCAI conferences of the last 10 years (2013-2022) are selected for review. Among the papers in these conferences (5 ICCV and 10 MICCAI), we select the ones having the word “segment” on their title. Then for each year, we have randomly selected 20 papers to review, having a total of 300 papers.

**General Notes**

Each paper will be reviewed by 2 reviewers, and if there are discrepancies, they will be discussed by the two reviewers and if needed with others. Therefore, in total we will have 600 reviews to do.

An Excel file named “Review” with 600 rows and the questions will be shared. And each reviewer will also receive a personal excel file containing only their assigned papers and some empty rows. Each reviewer will do the review in their personal file, and we will discuss them after the entire review is completed.

The first 150 rows of the shared “Review” file have the corresponding reviewer initials on the “Reviewer” column. Each reviewer must review at least 30 papers, so since we are 5 the first 150 rows are already assigned to the reviewers, and they are also copied to each reviewer’s personal file. For the rest of the papers, any reviewer who is ready and willing to review the next paper will first put their initial on the “Reviewer” column of the shared file and then copy the entire row to the empty rows of their personal file to review. It is important to first put your initials in the shared file and then do the review, so others will know this paper is already being reviewed.

If a paper is not using machine learning, meaning that it does not have a training and testing procedure, we will exclude that paper. So, for such a paper just put your ID with an X mark (e.g. RW1-X) and skip to the next paper.

**Questions**

The following questions are considered for the review. To answer these questions, we first will read only the experiments section of the paper and try to answer the questions. If the answer is not clear from the text, we can do extra keyword searching in the entire paper that are suggested for each question in the table below. If the answer is still unclear, we would select the “Not clear/Insufficiently described” option.

|  |
| --- |
| The blue questions (1-3) are about dataset. |
| 1- Is the size of the dataset stated clearly? |
| If the exact size of the used datasets are both stated and referenced. We would answer: “Yes, stated or possible to calculate”  If Some of the datasets are referenced and some stated we would answer: “Yes, stated or possible to calculate”  **Keywords:** dataset, data, train, test  **Possible original answers and their mappings to the final answer options presented in the paper:**  Yes, stated or possible to calculate → Yes  Referenced → Yes  Some are stated and some are referenced → Yes  No → No  Not clearly stated → Note Clear |
| 2- Do they test on any public data? |
| We will not check if the data is accessible or how hard it is to access the data. If it is referenced or mentioned that it is available, we will answer yes.  **Keywords:** dataset, data, public  **Possible original answers and their mappings to the final answer options presented in the paper:**  Yes → Yes  No → No |
| 3- Do they consider stratification? |
| **Keywords**: stratif, balanc, ratio  By stratification we mean that they are dividing a dataset into **any** subsets or subgroups based on **any** criteria or characteristics. Or they are considering or talking about the subgroups, or minority groups in their data. Even if they do not actually perform it, but they just consider it or talk about it we would answer: Considered.  These are examples of considering stratification:  1- They are preserving the same ratio of each class in train and test set.  2- They are including same ratio of adults and children in their data.  3- Our method may not perform well for pregnant people, since we have a small number of them in the data.  4- Other possible variables to stratify on include smoking habits, age, sex, etc.  **Possible original answers and their mappings to the final answer options presented in the paper:**  Considered → Yes  Not considered → No  Not clear/Insufficiently described → Not Clear |
| The green questions (4-10) are about information leakage. Each question is evaluating the risk of information leakage in the corresponding task. Finally question 10 must be answered based on the previous ones to conclude about the independency of the test data. |
| 4- Preprocessing? |
| **Keywords**: preprocess, normalize, scale, …  If they are performing pre-processing on entire data before split but they are not introducing any dependencies between train and test data we will answer: “Preprocessing the entire data before split with no risk of leakage”  **Possible original answers and their mappings to the final answer options presented in the paper:**  No preprocessing → Not Applicable  Preprocessing separately after splitting → Yes  Preprocessing the entire data before split with no risk of leakage → Yes  Preprocessing the entire data before split with risk of leakage → No  Not clear → Not Clear |
| 5- Data augmentation? |
| **Keywords**: augment  **Possible original answers and their mappings to the final answer options presented in the paper:**  No augmentation → Not Applicable  Done on train data. → Yes  Done on the entire data before split. → No  Not clear/Insufficiently described → Not Clear |
| 6- Data over/under-sampling? |
| **K eywords**: over-sampling, imbalance  Over/under-sampling are techniques used to address class imbalance in a dataset. Over-sampling involves creating new synthetic data for the minority class, and under-sampling involves decreasing the number of samples in the majority class.  Over-sampling the entire data can introduce leakage (similar to data augmentation)  Under-sampling can also introduce leakage in some cases if the method they are using is utilizing neighborhood information.  **Possible original answers and their mappings to the final answer options presented in the paper:**  No over-sampling → Not Applicable  Done on train data. → Yes  Done on entire data before split. → No  Not clear → Not Clear |
| 7- 3D data with 2D models? |
| **Keywords**: video, sequence, temporal, 3D, slice, volume  We are looking for papers who have 3D data (like video sequences or 3D images such as volumetric images) and feed it into 2D models. That is, they use 2D slices of the data (video frames, MRI slices, …) to train the model. We want to know if during the split of the data into train and test sets they are considering the dependencies in the 3rd dimension.  **Possible original answers and their mappings to the final answer options presented in the paper:**  Not used → Not Applicable  There is risk for dependence between train and test data → No  There is no risk for dependence between train and test data → Yes  Not clear/Insufficiently described → Not Clear |
| 8- Hyper-parameter tuning? |
| **Keywords**: hyper-parameter, parameter, tune, tuning  If they mention the values of parameters (es: lr=0.03) -> Not used  **Possible original answers and their mappings to the final answer options presented in the paper:**  Not used → Not Applicable  Done on train/validation set → Yes  Test set is included → No  Not clear/Insufficiently described → Not Clear |
| 9- Re-sampling? |
| **Keywords**: hyper-parameter, parameter, tune, tuning  Two of commonly used re-sampling techniques include **cross-validation** and **bootstrap sampling**. They usually create multiple subsets of the original data and use them in an iterative process for training and validating the model. If any of the previously mentioned steps (pre-processing, data augmentation, and over-sampling) are performed outside the cross-validation loop, information from the validation fold may leak into the training process.  **Possible original answers and their mappings to the final answer options presented in the paper:**  Not used → Not Applicable  Done correctly within the loop → Yes  Done incorrectly outside the loop → No  Not clear/Insufficiently described → Not Clear |
| 10- Is the test set independent? |
| We will answer this question based on the answers for questions 4-9.  **Possible original answers and their mappings to the final answer options presented in the paper:**  Yes: If the answers to all the questions (4-9) indicate the independency of the test set. → Yes  No: If the answer to at least one of the questions (4-9) indicates the dependency of test set → No  Not clear: If the answer to at least one of the questions (4-9) is Not clear/ Insufficiently described. → Not Clear |
| Questions about evaluation and report |
| 11- Are they fine tuning the baseline models? |
| **Keywords**: hyper-parameter, parameter, tune, tuning  If just mention the baseline but without saying nothing more, we will say -> NO  **Possible original answers and their mappings to the final answer options presented in the paper:**  No baselines/Nothing to tune → Not Applicable  Yes thoroughly. → Yes  Yes partially. → Yes  No → No  Not clear/Insufficiently described → Not Clear |
| 12- Are they reporting uncertainties? |
| **Keywords**: standard, mean, p-value, statistical test, …  **Possible original answers and their mappings to the final answer options presented in the paper:**  Yes → Yes  No → No  Not clear/Insufficiently described → Not Clear |
| 12.5- On which data they compute uncertainties? |
| To clarify the options, note that **train**, **validation**, and **test** sets are all subsets of one dataset:  train set: is used for training the model  validation set: is used for monitoring the train process and hyperparameter tuning  test set: for test and evaluating the trained model  and by **independent test** set we mean a separate dataset which is collected from another source, location, … and it is used for evaluating the model, so it is not a subset of the data used for training.    **Possible original answers and their mappings to the final answer options presented in the paper:**  Validation set  Cross-validation or other resampling methods  Test set  Independent test set  Other  Not clear/Insufficiently described |
| 13- Do they share codes? |
| **Keywords**: code, github, share  **Possible original answers and their mappings to the final answer options presented in the paper:**  Yes → Yes  No → No |