Corresponding author(s): DBPR
Last updated by author(s): September 25, 2024

Machine Learning Reporting Summary

1. Availability and reproducibility of Code and Data

Please select all that apply regarding the availability of the data and code used in the study.

☑ Code will be included in a CodeOcean capsule.
☑ The source code is included in the submission or available in a public repository: https://github.com/pikapi-pi/HeMiCoRe (url)
☐ A compiled standalone version of the software is included in the submission or available in a public repository: _____ (url)
☑ A test dataset and instructions/scripts for replicating the results are included in the submission or available in a public repository: https://github.com/pikapi-pi/HeMiCoRe (url)
☑ A Readme file with instructions for installing and running the code is included in the submission or available in a public repository: https://github.com/pikapi-pi/HeMiCoRe (url)
☑ The code is made available to reviewers during review.
☑ Pretrained models are used in the study and accessible through: https://download.pytorch.org/models/resnet18-5c1 (url)
☐ Pretrained models are used in the study and are not accessible.
☑ The paper contains information on how to obtain code and data after publication.

2. Datasets

A. All data sources are listed in the paper.
   ☐ Yes
   ☑ No (please justify) We have used the TCGA data and some private cohorts for the model training and validation. Since we opted for the double-blind peer review (DBPR), the sources of
B. The train, test and validation datasets are publicly available and links/accession numbers have been provided in the manuscript or supplementary materials.
   ☑ Yes
   ☐ No (please justify)_____
C. We have reported and discussed potential dataset biases in the paper. Where applicable, appropriate mitigation strategies were used.
   ☑ Yes (please specify section) Please see them in the section Results
   ☐ No (please justify) _____

D. The data cleaning and preprocessing steps are clearly and fully described, either in text or as a code pipeline.
   ☑ Yes (please specify section) Please see it in the readme.md at https://github.com/pikapi-pi/HeMiCoRe
   ☐ No (please justify) _____
E. Instances of combining data from multiple sources are clearly identified, and potential issues mitigated.
   ☑ Yes (please specify section) We have conducted some cross-site validation experiments to investigate the effect of different data sources. Please see them in the section Results.
   ☐ No (please justify) _____

3. Model and training

A. What model architecture is the current model based on? Hypergraph Neural Network
B. A Model Card is provided[1].
   ☑ Yes
   ☐ No
C. The model clearly splits data into different sets for training (model selection), validation (hyperparameter optimization), and testing (final evaluation).
   ☑ Yes
   ☐ No
D. The method of data splitting (e.g. random, cluster- or time-based splitting, forward cross-validation) is clearly stated.
   ☑ Yes (please specify) We used five-fold stratified cross-validation
   ☐ No (please justify) _____
E. The data splitting mimics anticipated real-world applications.
   ☐ Yes
   ☑ No
F. The data splitting procedure has been chosen to avoid data leakage.
   ☑ Yes (please specify) The data splitting procedure has been chosen to avoid data leakage and keep the balance of samples.
   ☐ No (please justify) _____
G. The interpretability of the model has been studied and clearly validated.
   ☑ Yes (please specify section) Please see it in the section Results.
   ☐ No (please justify) _____

4. Evaluation

A. The performance metrics used are described and justified in the paper.
   ☑ Yes (please specify section) Please see it in the section Results and Supplementary.
   ☐ No (please justify) _____

[1] https://huggingface.co/docs/hub/model-cards

B. Cross-validation of the results is included.
   ☑ Yes
   ☐ No
C. Community-accepted benchmark datasets/tasks are used for comparisons.
   ☑ Yes (please specify) The majority of our datasets are from the community-accepted public data source TCGA. The benchmark task is survival prediction.
   ☐ No
D. Baseline comparisons to simple/trivial models (for example, 1-nearest neighbour, random forest, most frequent class) are provided.
   ☐ Yes
   ☑ No
E. Benchmarks with current state-of-the-art are provided.
   ☑ Yes
   ☐ No
F. Ablation experiments are included.
   ☐ Yes
   ☑ No
G. The model has been tested on a fully independent dataset.
   ☑ Yes
   ☐ No

5. Computational resources

A. The paper contains information on hardware/computing resources that were used.
   ☑ Yes
   ☐ No
B. The paper includes information on the computational costs in terms of computation time, parallelization or carbon footprints estimates.
   ☑ Yes
   ☐ No