Author Contributions Checklist Form

This form documents the artifacts associated with the article (i.e., the data and code supporting the computational findings) and describes how to reproduce the findings.

Part 1: Data

☐ This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Abstract

The Physionet 2012 Challenge dataset is an EHR dataset that contains 12,000 ICU patients with measurements of 37 variables across 48 time points, with one hour between each consecutive time point. Preprocessing of the data was done using the code in this repository, which uses prior domain knowledge to compress the repeated measurements into one observation per patient. The patients are divided into training, validation, and test sets of equal sizes, and an entry is missing if, for a patient, the corresponding feature was not measured at any time point.

Availability

☒ Data **are** publicly available.
☐ Data **cannot be made** publicly available.

If the data are publicly available, see the *Publicly available data* section. Otherwise, see the *Non-publicly available data* section, below.

Publicly available data

☒ Data are available online at: Various locations. See the supplementary materials of the NIMIWAE manuscript (NIMIWAE_Appendix.pdf) for details on these datasets. Missingness is simulated on top of the UCI datasets, while missingness is inherent in the Physionet dataset. The missingness-simulated datasets can be reproduced by using the code in the reproducibility repository (using the default seed as provided), and these datasets can also be provided by request. These links are copied below for your convenience.

banknote Documentation of the data can be found here. The data can be found here

concrete Documentation of the data can be found here. The data can be found here

hepmass Documentation of the data can be found here. The data can be found here

power Documentation of the data can be found here. The data can be found here

red Documentation of the data can be found here. The data can be found here

white Documentation of the data can be found here. The data can be found here

Physionet 2012 Challenge data The raw data can be downloaded here. The pre-processed version of the data can be found here.

☐ Data are available as part of the paper's supplementary material.

☐ Data are publicly available by request, following the process described here:

☐ Data are or will be made available through some other mechanism, described here:

Description

File format(s)

☒ CSV or other plain text.
☐ Software-specific binary format (.Rda, Python pickle, etc.):
☐ Standardized binary format (e.g., netCDF, HDF5, etc.):
☐ Other (please specify):

Data dictionary

☒ Provided by authors in the following file(s): See the Appendix of the NIMIWAE manuscript (NIMIWAE_Appendix.pdf)

☐ Data file(s) is(are) self-describing (e.g., netCDF files)

☐ Available at the following URL:

Additional Information (optional)

Part 2: Code

Abstract

This code reproduces training of models and summarization of results in the paper "Handling Non-ignorably Missing Features in Electronic Health Records Data Using Importance-Weighted Autoencoders." We show superiority of NIMIWAE in imputation of simulated missing values, and perform comparative analyses on the Physionet 2012 Challenge EHR Dataset, showing improvements in imputed missing entries.

Description

Code format(s)

- ☒ Script files
  - ☒ R
  - ☒ Python
  - ☐ Matlab
  - ☐ Other:
- ☒ Package
  - ☒ R
  - ☒ Python
  - ☐ MATLAB toolbox
  - ☐ Other:
- ☒ Reproducible report
  - ☒ R Markdown
  - ☐ Jupyter notebook
  - ☐ Other:
- ☐ Shell script
- ☐ Other (please specify):

Supporting software requirements

Version of primary software used

- R version 3.6.1
- Python version 3.6.3

Libraries and dependencies used by the code R packages:

- reticulate (1.13)
- NIMIWAE (0.1.0)

Python modules:

- numpy (1.18.1)
- pandas (1.5.0)
- scipy (1.4.1)
- torch (1.5.0)
- tensorflow (1.14.0)
- sklearn (0.22.1)
- argparse (1.1)
- tqdm (4.42.1)

Supporting system/hardware requirements (optional)

This code requires access to a cuda-enabled GPU.

Parallelization used

☒ No parallel code used
☐ Multi-core parallelization on a single machine/node

- Number of cores used:

☐ Multi-machine/multi-node parallelization

- Number of nodes and cores used: 3 nodes, 243 cores

License

☒ MIT License (default)
☐ BSD
☐ GPL v3.0
☐ Creative Commons
☐ Other: (please specify below)

Additional information (optional)

The blinded NIMIWAE package can be found here. It has also been attached in the original submission for convenience.

Part 3: Reproducibility workflow

Scope

The provided workflow reproduces:

☒ Any numbers provided in text in the paper
☒ All tables and figures in the paper
☐ Selected tables and figures in the paper, as explained and justified below:

Workflow

Format(s)

☐ Single master code file
☐ Wrapper (shell) script(s)
☐ Self-contained R Markdown file, Jupyter notebook, or other literate programming approach
☐ Text file (e.g., a readme-style file) that documents workflow
☐ Makefile
☒ Other (more detail in *Instructions* below)

Instructions

Run `runComparisons.R` script to train all models (time-consuming). Then, run `SummarizeResults.R` script to reproduce Figures 2 and 3, and the results in Table 2.

Expected run-time

Approximate time needed to reproduce the analyses on a standard desktop machine:

☐ < 1 minute
☐ 1-10 minutes
☐ 10-60 minutes
☐ 1-8 hours
☒ > 8 hours
☐ Not feasible to run on a desktop machine, as described here

Training the deep learning architectures can be very time-consuming, and it may be recommended to run on a computing cluster with access to a GPU with a large memory (at least 16GB recommended). Summarizing the results should not take any longer than 10 minutes.

Additional information (optional)

The blinded reproducibility repository can be found here, and it can also be found attached in the original submission for convenience.

Notes (optional)