## Overview

The code in this replication package constructs the analysis file from the public database, TCIA, the cohort of NSCLC Radiogenomics, using Python.

## Data Availability and Provenance Statements

* ☐ 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).

### Statement about Rights

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

### Summary of Availability

* √ All data **are** publicly available.
* ☐ Some data **cannot be made** publicly available.
* ☐ **No data can be made** publicly available.

Data were downloaded from The Cancer Imaging Archice (TCIA), organized by National Cancer Institute (NIH). We use the collection of “NSCLC-Radiogenomics”. Data can be downloaded from https://wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics. (Bakr S, Gevaert O, Echegaray S, et al. A radiogenomic dataset of non-small cell lung cancer. Sci Data. 2018;5:180202.)

A copy of the data is provided as part of this archive. The data are in the public domain.

## Dataset list

|  |  |  |  |
| --- | --- | --- | --- |
| Data file | Source | Notes | Provided |
| data/feature.csv | Processed TCIA data | Image features extracted from CT images | Yes |
| data/gene.csv | TCIA | Gene expression profile | Yes |
| data/clinical data.xlsx | TCIA | Clinical data for patients used in the paper | Yes |
| data/spearman\_plot.csv | Processed data | Spearman analysis result, serves as input for Figure 1. | Yes |
| data/datadistribution\_plot.txt | Processed data | Data distribution, serves as input for Figure 4 | Yes |
| data/train.txt | Processed data | Training dataset serve as input for machine learning model | Yes |
| data/test.txt | Processed data | Test dataset serve as input for machine learning model | Yes |
| data/km.csv | Processed data | Predicted results of HOPX expression status from test for Kaplan-Meier analysis | Yes |

## Computational requirements

### Software Requirements

* Python 3.8.1
  + PyRadiomics 3.0.1
  + Lifeline 0.27.0
  + ML-ENS 0.2.3
  + Bayesian-optimization 0.6.0
  + Other package requirements could be seen on the guideline website of PyRadiomics, Lifeline, ML-ENS and Bayesian-optimization

### Memory and Runtime Requirements

#### Summary

Approximate time needed to reproduce the analyses on a standard (CURRENT YEAR) desktop machine:

* ☐ <10 minutes
* ☐ 10-60 minutes
* √ 1-8 hours
* ☐ 8-24 hours
* ☐ 1-3 days
* ☐ 3-14 days
* ☐ > 14 days
* ☐ Not feasible to run on a desktop machine, as described below.

#### Details

The code was last run on a **4-core Intel-based laptop with MacOS version 11.4.0**

## Description of programs/code

* Program in programs/01\_feature extraction will extract image features from CT images with ROIs. The file programs/01\_feature extraction should be run in the terminal of your computer.
* Programs in programs/02\_machinelearning generate the radiogenomic signature candidates and use this as input for machine model to predict survival and HOPX expression status. The program programs/02\_ machinelearning/normalization.ipynb will do the normalization to image features and gene expression. The program programs/02\_ machinelearning/spearman.py will calculate the spearman coefficients of all image features and select the pairs which coefficient is over 0.5. The program programs/02\_ machinelearning/BayesianOptimization.ipynb will select the best parameters for each base machine learning model (SVM, RF, and GBDT). The program programs/02\_ machinelearning/stacking.ipynb is the code for the ensemble learning machine model, stacking and will calculate the accuracy, sensitivity, specificity, ROC and AUC for model evaluation.
* Programs in programs/03\_plot will generate some of the figures in the paper. The program programs/03\_plot/heatmap.ipynb will generate the heatmap of Figure 1. The program programs/03\_plot/ Kaplan-Meier.ipynb will generate the Kaplan-Meier curves of predicted results of HOPX expression status from test for Figure 2 and Figure 6(b)and calculate the p-value (log-rank test) between two groups. The program programs/03\_plot/ data distribution.ipynb will generate the violin plot of Figure 4 and calculate the p-value (t test) between two groups of image features. The program programs/03\_plot/DEA\_volcano.ipynb will generate the volcano plot of Figure 6(a).

## List of tables and programs

The provided code reproduces:

* ☐ All 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Figure/Table # | Program | Line Number | Output file | Note |
| Figure 1 | 03\_plot/heatmap.ipynb |  | HeatMap.png |  |
| Figure 2 and Figure 6(b) | 03\_plot/ Kaplan-Meier.ipynb |  | KMoriginal.png  KM2.png  KM3.png  KM4.png  KM8.png |  |
| Figure 4 | 03\_plot/ data distribution.ipynb |  | original\_ske\_violin 2.png  wavelet\_root\_violin 2.png |  |
| Figure 6(a) | 03\_plot/DEA\_volcano.ipynb |  | dea.png |  |

## References

Bakr S, Gevaert O, Echegaray S, et al. A radiogenomic dataset of non-small cell lung cancer. Sci Data. 2018;5:180202.

Davidson-Pilon, Cameron, lifelines, survival analysis in Python, (2022). https://doi.org/10.5281/zenodo.6359609

Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillon-Robin, J. C., Pieper, S., Aerts, H. J. W. L. (2017). Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research, 77(21), e104–e107.

@misc{flennerhag:2017mlens,

author = {Flennerhag, Sebastian},

title = {ML-Ensemble},

month = nov,

year = 2017,

doi = {10.5281/zenodo.1042144},

url = {https://dx.doi.org/10.5281/zenodo.1042144}

}

## Acknowledgements

Some content on this page was copied from [Hindawi](https://www.hindawi.com/research.data/#statement.templates). Other content was adapted from [Fort (2016)](https://doi.org/10.1093/restud/rdw057), Supplementary data, with the author’s permission.