| **Disease** | **Description** | **Source** | **Imputation Strategy** | **Preprocessing** | **Access Link** |
| --- | --- | --- | --- | --- | --- |
| Chronic kidney disease | Clinical and laboratory data | Rubini, Soundarapandian  , Eswaran  et al. | Mean |  | https://doi.org/10.24432/C5G020 |
| Diabetes | Clinical and laboratory data | Vanschoren, van Rijn, Bischl, Torgo et al. |  |  | https://www.openml.org/search?type=data&status=active&id=37&sort=runs |
| Ovarian  cancer | Clinical and laboratory data | Lu, Fan et al. | Mean | Removing illegal characters | https://doi.org/10.1016/j.ijmedinf.2020.104195 |
| Sepsis | Clinical and laboratory data | Reyna et al. | Mean | Removing samples with missing values | https://doi.org/10.1097/CCM.0000000000004145 |
| Health and nutrition survey | Clinical and laboratory data | cdc.gov | K-Nearest Neighbors | Random sampling | https://wwwn.cdc.gov/nchs/nhanes/Default.aspx |
| Smoking status | Clinical and laboratory data | data.go.kr |  | Random sampling | https://github.com/Koda98/smoker-status-prediction?tab=readme-ov-file |
| Tumors | RNA sequencing data | O’Neill et al. |  | Quality control | https://doi.org/10.1016/j.xgen.2024.100674 |
| Breast  cancer | Single-cell sequencing data | Asemota et al. |  | Quality control | https://doi.org/10.1073/pnas.2406837121 |
| Lung cancer | Single-cell sequencing data | Yang, Xu, Zhou, Liu et al. |  | Quality control | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE245170 |
| Diabetes | Single-cell sequencing data | Qian et al. |  | Quality control | https://doi.org/10.1186/s12933-024-02458-x |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Disease** | **Type** | **Attributes After Preprocessing** | **Missing data imputation** | **Link** |
| Chronic Kidney Disease | Clinical & serum markers | Shape:  400 obs × 25 features |  | https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease |
| Diabetes | Clinical & serum markers | Shape:  768 obs × 9 features |  | https://www.openml.org/search?type=data&status=active&id=37&sort=runs |
| Ovarian cancer | Clinical & serum markers | Shape:  349 obs × 50 features |  | https://doi.org/10.1016/j.ijmedinf.2020.104195 |
| Sepsis | Clinical & serum markers | Shape:  822 obs × 42 features |  | https://www.kaggle.com/datasets/salikhussaini49/prediction-of-sepsis |
| Smoking-related | Clinical & serum markers | Shape:  11695 obs × 23 features |  | https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction-using-biosignals/data |
| National health and nutrition-examination survey lab | Clinical & serum markers | Shape:  9813 obs × 47 genes |  | https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey |
| Cancers | RNA seq | Shape:  187 obs × 500 genes |  | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE270257 |
| Breast cancer | Single-cell RNA seq | Shape:  4018 rows × 300 genes |  | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE274729 |
| Lung cancer tracheal adenoid cystic carcinoma | Single-cell RNA seq | Shape:  552 obs × 300 genes |  | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE245170 |
| Diabetes | Single-cell RNA seq | Shape:  961 obs × 300 genes |  | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE248608 |
|  |  |  |  |  |
|  |  |  |  |  |

| **Disease** | **Type** | **Description** | **Access Information** | **License** | **Access Link** |
| --- | --- | --- | --- | --- | --- |
| MNIST | Yann LeCun | Handwritten digits dataset | [MNIST Website](http://yann.lecun.com/exdb/mnist/) | CC BY-SA 3.0 | Training |
| CIFAR-10 | Alex Krizhevsky | Small image classification | [CIFAR-10 Website](https://www.cs.toronto.edu/~kriz/cifar.html) | MIT License | Training |
| IMDB Reviews | Stanford AI Lab | Sentiment analysis dataset | [IMDB Dataset](https://ai.stanford.edu/~amaas/data/sentiment/) | Custom License | Analysis |