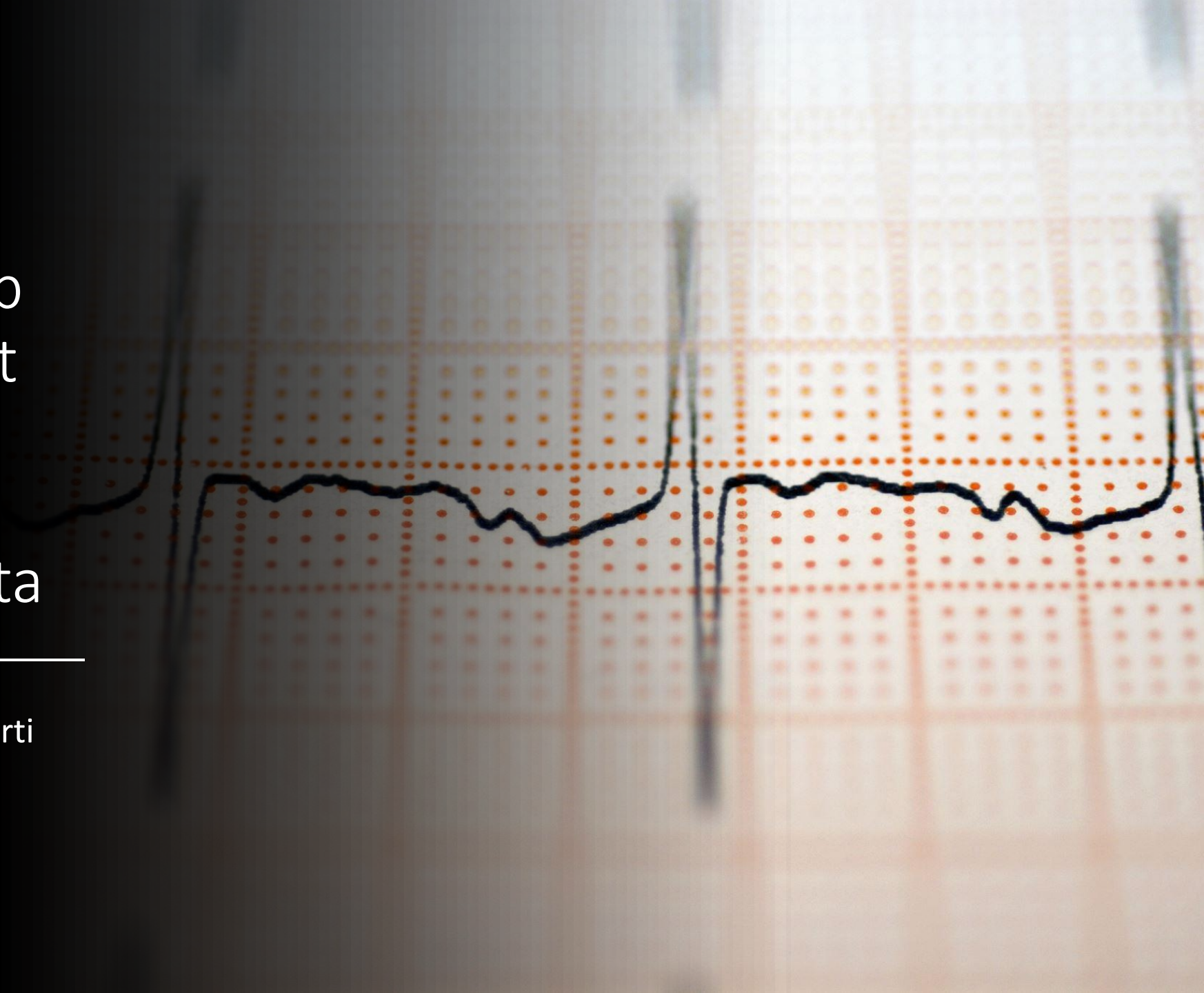
Reproducing Deep Learning for Heart Failure Onset Prediction Using Temporal EHR Data

A Reproduction Study - Shiv Bharti



Importance of Early Detection

Heart failure affects millions worldwide

Early detection is crucial for effective treatment

EHRs provide a rich data source for predictive models

Objective of the Original Paper

- Explore the use of deep learning for modeling temporal relations among events in EHRs
- Compare the performance of RNN models to conventional methods for predicting heart failure onset

Scope of Reproducibility

- Re-implement the original study's methodology
- Validate results by comparing RNN models to conventional methods
- Focus on 12- and 18-month observation windows

Reproduction Results

- **12-month observation window**
- RNN AUC: 0.714
- Logistic Regression AUC: 0.788
- MLP AUC: 0.762
- SVM AUC: 0.850
- KNN AUC: 0.703

- **18-month observation window**
- RNN AUC: 0.784
- Logistic Regression AUC: 0.788
- MLP AUC: 0.754
- SVM AUC: 0.850
- KNN AUC: 0.703

Automated ML - Microsoft Azur

https://ml.azure.com/automl/welcome?wsid=/subscriptions/46f9e149-9912-4123-a410-773f7034294b/resourcegroups/Product-Demo/providers/Microsoft.MachineLearningServices/workspaces/shivbhart...

Microsoft Azure Machine Learning Studio

shivbhart
shivbhart

Microsoft Non-Production > shivbhart > Automated ML

Automated ML

Let Automated ML train and find the best model based on your data without writing a single line of code. Learn more about Automated ML

- All workspaces
- Home
- Model catalog PREVIEW

Authoring

- Notebooks
- Automated ML
- Designer

Assets

- Data
- Jobs
- Components
- Pipelines
- Environments
- Models
- Endpoints

Manage

- Compute
- Linked Services
- Data Labeling

+ New Automated ML job Refresh

Recent Automated ML jobs

View all experiments

Display name	Experiment	Status	Created on	Duration	Created by	Compute target	Tags
DL_HF_Prediction_Reproduction_Study_	HF_Prediction_GPU	Completed	May 3, 2023 10:07 AM	53m 33s	Shiv Bharti	ml-compute-sb	dynamic_allowlisti
DL_HF_Prediction_Reproduction_Study_	HF_prediction	Completed	May 3, 2023 8:21 AM	1h 23m 42s	Shiv Bharti	ml-compute-sb	dynamic_allowlisti

Docum

Compute - Microsoft Azure Ma

https://ml.azure.com/compute/list?wsid=/subscriptions/46f9e149-9912-4123-a410-773f7034294b/resourcegroups/Product-Demo/providers/Microsoft.MachineLearningServices/workspaces/shivbharti&ti...

Microsoft Azure Machine Learning Studio

shivbhart
shivbhart

Microsoft Non-Production > shivbhart > Compute

Compute

The "Kubernetes clusters" tab is now where you can access previous versions of "inference clusters" (also known as "AKS clusters") and "attached Kubernetes" compute types along with any previously created compute targets using those types. Learn more about Kubernetes clusters.

Compute instances | Compute clusters | Kubernetes clusters | Attached computes

+ New Refresh Start Stop Restart Schedule and idle shutdown Delete View options View quota

Search Filter

Name	State	Idle shutdown	Applications	Size	Created on	Assigned to
sbharti2	Stopped	15 minutes	JupyterLab Jupyter VS Code (Desktop) PREVIEW	STANDARD_E4DS_V4	May 3, 2023 9:59 AM	Shiv Bharti
sbharti1	Stopped	--	JupyterLab Jupyter VS Code (Desktop) PREVIEW	STANDARD_DS12_V2	Mar 20, 2023 5:59 PM	Shiv Bharti

https://ml.azure.com/automl/welcome?wsid=/subscriptions

74°F
Partly sunny

2:47 PM
5/5/2023

Notebooks - Microsoft Azure M

https://ml.azure.com/fileexplorerAzNB?wsid=/subscriptions/46f9e149-9912-4123-a410-773f7034294b/resourcegroups/Product-Demo/providers/Microsoft.MachineLearningServices/workspaces/shivbhart...

Microsoft Azure Machine Learning Studio

shivbharti
shivbharti

Microsoft Non-Production > shivbharti > Notebooks

hf_prediction_using_F

All workspaces

Home

Model catalog PREVIEW

Authoring

Notebooks

Automated ML

Designer

Assets

Data

Jobs

Components

Pipelines

Environments

Models

Endpoints

Manage

Compute

Linked Services

Data Labeling

Files

Edit in VS Code | Compute: sbharti2 - Stopped | No kernel connected

Viewing | Last saved a few seconds ago

```python
import os
import numpy as np
import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import DataLoader, Dataset
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.neural_network import MLPClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import roc_auc_score
from sklearn.model_selection import train_test_split
import random
import pickle

# Set seed
seed = 24
random.seed(seed)
np.random.seed(seed)
torch.manual_seed(seed)
os.environ["PYTHONHASHSEED"] = str(seed)

# Define data path
DATA_PATH = "adl://sbharti.azuredatalakestore.net/path/to/DL_HC_RNN-lib/data"

# Load data
pids = pickle.load(open(os.path.join(DATA_PATH, 'train/pids.pkl'), 'rb'))
vids = pickle.load(open(os.path.join(DATA_PATH, 'train/vids.pkl'), 'rb'))
hfs = pickle.load(open(os.path.join(DATA_PATH, 'train/hfs.pkl'), 'rb'))
seqs = pickle.load(open(os.path.join(DATA_PATH, 'train/seqs.pkl'), 'rb'))
types = pickle.load(open(os.path.join(DATA_PATH, 'train/types.pkl'), 'rb'))
rtypes = pickle.load(open(os.path.join(DATA_PATH, 'train/rtypes.pkl'), 'rb'))

input_size = len(types)

def pad_and_encode_sequence(sequence, max_visits, max_seq_length):
    padded_sequence = np.zeros((max_visits, max_seq_length))
    for i, visit in enumerate(sequence[:max_visits]):
```

74°F
Partly sunny

Search

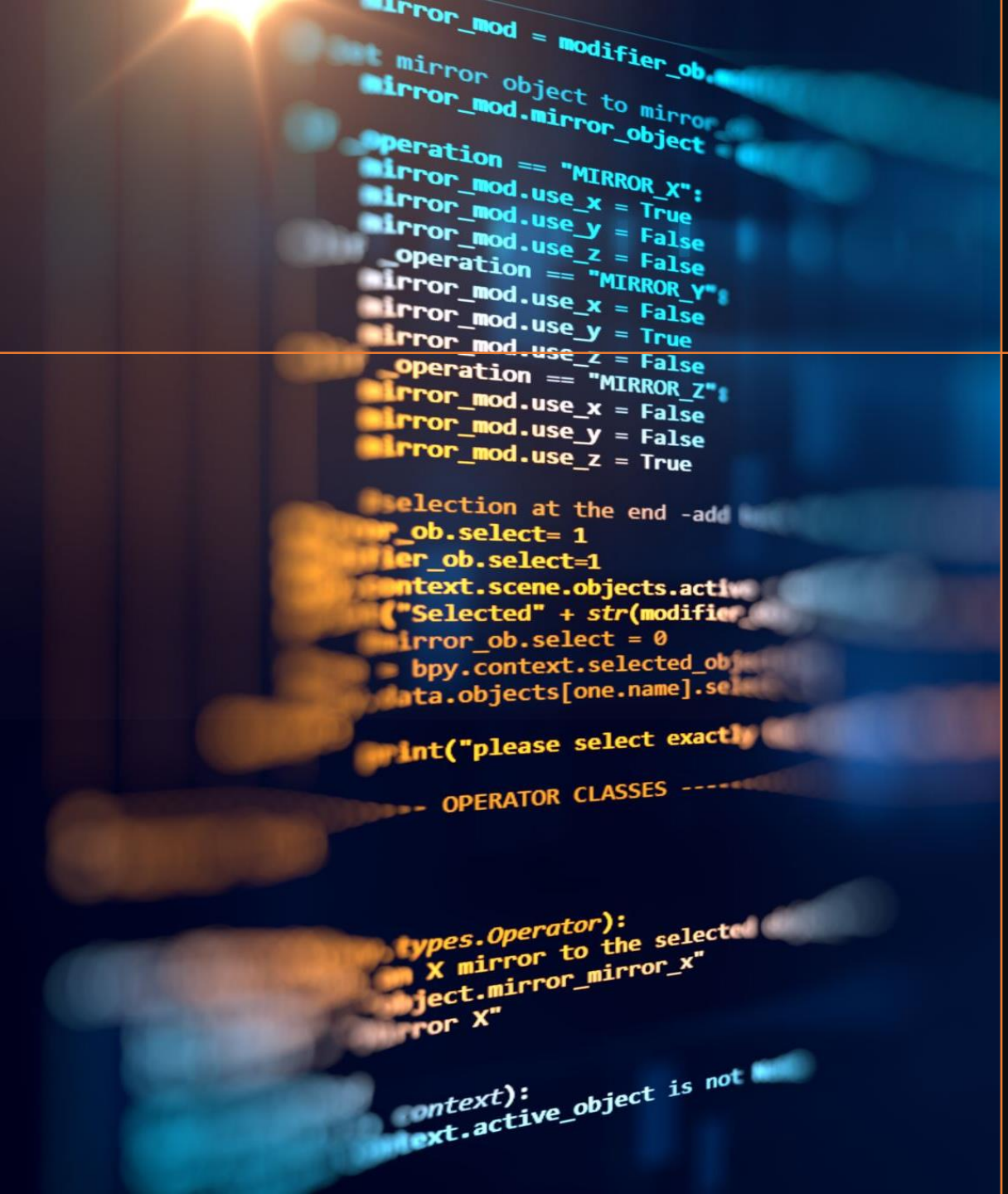
2:49 PM
5/5/2023

Results

```
print(f"Logistic Regression AUC: {aucs_12[1]:.3f}")
print(f"MLP AUC: {aucs_12[2]:.3f}")
print(f"SVM AUC: {aucs_12[3]:.3f}")
print(f"KNN AUC: {aucs_12[4]:.3f}")

aucs_18 = train_and_evaluate_models(train_seqs_18, test_seqs_18, train_hfs_18, test_hfs_18)
print("Results for 18-month observation window:")
print(f"RNN AUC: {aucs_18[0]:.3f}")
print(f"Logistic Regression AUC: {aucs_18[1]:.3f}")
print(f"MLP AUC: {aucs_18[2]:.3f}")
print(f"SVM AUC: {aucs_18[3]:.3f}")
print(f"KNN AUC: {aucs_18[4]:.3f}")
```

```
Results for 12-month observation window:
RNN AUC: 0.714
Logistic Regression AUC: 0.788
MLP AUC: 0.762
SVM AUC: 0.850
KNN AUC: 0.703
Results for 18-month observation window:
RNN AUC: 0.784
Logistic Regression AUC: 0.788
MLP AUC: 0.754
SVM AUC: 0.850
KNN AUC: 0.703
```



Challenges and Recommendations

Challenges:

- Obtaining access to EHR data
- Limited data availability

Recommendations:

- Data sharing (de-identified or synthetic)
- Clear documentation
- Open-source code
- Benchmarking on public datasets

English

CITI PROGRAM

My Courses | My Records | My CE/CMEs | Support

Shiv Bharti
ID 12222711

Records

University of Illinois at Urbana - Champaign (ID 486)

Show Records for: University of Illinois at Urbana - Champaign | Show All

University of Illinois at Urbana - Champaign Records (ID 486)

Additional UIUC Training - Conflicts of Interest in Research Involving Human Subjects - Biomedical (ID 22758)

Stage	Record ID	Passing Score	Your Score	Start Date	Completion Date	Expiration Date	Gradebook	Completic Record
Basic Course	55704866	80%	100%	04-May-2023	04-May-2023	04-May-2026	View	View-Print-S

Additional UIUC Training - Research and HIPAA Privacy Protections (ID 22755)

Stage	Record ID	Passing Score	Your Score	Start Date	Completion Date	Expiration Date	Gradebook	Completic Record
Basic Course	55706548	80%	100%	04-May-2023	04-May-2023	04-May-2026	View	View-Print-S

Your application for PhysioNet training - Bharti, Shiv Ratan - Outlook - Google Chrome

about:blank

Your application for PhysioNet training

You forwarded this message on Thu 5/4/2023 2:36 PM

PS

PhysioNet Automated System <noreply@physionet.org>
To: Bharti, Shiv Ratan
Thu 5/4/2023 2:29 PM

Dear Shiv Bharti,

We are pleased to say that your "CITI Data or Specimens Only Research" training was approved.

You are now able to access protected databases upon agreeing to the terms of usage. For example, you can access MIMIC-III by following the steps below:

- Go to the project page at https://urldefense.com/v3/__https://physionet.org/content/mimiciii/__;!!DZ3fjg!6bi2u0QltfrxNPT9VJqRm6kyJ6LU5H87I2I862_0-veduppppcvbehqk90ofPLTYF7hYDisjxYsHTbjpRNjI-g$
- Find the "Files" section in the project description
- Click "Sign the data use agreement" to agree to the terms of usage for this dataset

Regards, The PhysioNet Team, MIT Laboratory for Computational Physiology Institute for Medical Engineering and Science, MIT, E25-505 77 Massachusetts Ave, Cambridge, MA 02139

physionet.org/content/mimiciii/1.4/

Bookmarks Life Personal Home Time To-Do Tax Market Glossary MS Ref Train & Certify

MIMIC-III Clinical Database

Alistair Johnson , **Tom Pollard** , **Roger Mark**

Published: Sept. 4, 2016. Version: 1.4

MIMIC-III available on the cloud *(Sept. 30, 2019, 2:28 p.m.)*

The MIMIC-III database is now available on two major cloud platforms: Google Cloud Platform (G
data on the cloud, simply add the relevant cloud identifier to your PhysioNet profile. Further inst

Tutorials are available for using MIMIC-III with BigQuery (GCP) and using MIMIC-III with Athena

When using this resource, please cite: (show more options)
Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). *PhysioNet*. https://doi.org/10.13026/C2XW26.

Additionally, please cite the original publication:
Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.

Please include the standard citation for PhysioNet: (show more options)
Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

Abstract

MIMIC-III is a large, freely-available database comprising deidentified health-related data associate
with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Med
Center between 2001 and 2012. The database includes information such as demographics, vital sigr
measurements made at the bedside (~1 data point per hour), laboratory test results, procedures,
medications, caregiver notes, imaging reports, and mortality (including post-hospital discharge).

physionet.org/content/mimiciii/1.4/

Bookmarks Life Personal Home Time To-Do Tax Market Glossary MS Ref Train &

Files

Total uncompressed size: 6.2 GB.

Access the files

- Download the ZIP file (6.2 GB)
- Request access to the files using the Google Cloud Storage Browser. Login with a Google account is required.
- Access using AWS Open Data repository
- Request access to the data using AWS command line tools: `aws s3 sync s3://mimic-iii-physionet DESTINATION`
- Request access using Google BigQuery.
- Download the files using your terminal: `wget -r -N -c -np --user shivbharti --ask-password https://physionet.org/files/mimiciii/1.4/`

Folder Navigation: <base>

Name	Size	Modified
ADMISSIONS.csv.gz	2.4 MB	2019-03-19
CALLOUT.csv.gz	1.1 MB	2019-03-19
CAREGIVERS.csv.gz	48.4 KB	2019-03-19
CHARTEVENTS.csv.gz	4.0 GB	2019-03-19
CPTEVENTS.csv.gz	4.7 MB	2019-03-19
DATETIMEEVENTS.csv.gz	52.5 MB	2019-03-19
DIAGNOSES_ICD.csv.gz	4.5 MB	2019-03-19
DRGCODES.csv.gz	1.7 MB	2019-03-19
D_CPT.csv.gz	3.9 KB	2019-03-19
D_ICD_DIAGNOSES.csv.gz	278.3 KB	2019-03-19
D_ICD_PROCEDURES.csv.gz	74.1 KB	2019-03-19
D_ITEMS.csv.gz	183.5 KB	2019-03-19
D_LABITEMS.csv.gz	11.2 KB	2019-03-19
ICUSTAYS.csv.gz	1.9 MB	2019-03-19
INPUTEVENTS_CV.csv.gz	402.6 MB	2019-03-19
INPUTEVENTS_MV.csv.gz	143.9 MB	2019-03-19
LABEVENTS.csv.gz	320.3 MB	2019-03-19

Conclusion

- Reproduced the original study with some variations in results
- RNN models showed potential for heart failure prediction
- Challenges in data access and limited data availability