The information of each folder and their files are introduced as follows:

# RQ1: Deep representation VS Shallow metrics

1. Acronyms and Abbreviations

LR Logistic Regression

BN Bayesian Network

RF Random Forest

dr Deep Representations

sm Shallow Metrics

1. File Content Description
2. i.mat (i=1,2,..,50) A standard data format file for MATLAB, it’s a cell array which has five elements. Each element is a 2\*2 confusion matrix, [TN,FP; FN,TP]. i-th confusion matrix corresponds to the results of i-th fold run (5 folds cross validation (CV) is used in our paper). The average result of 5 folds CV is regarded as the result of one run.
3. Models.mat It’s a cell array including two cell elements, and each element is a 50\*7 array where 50 means the total runs, 7 denotes the performance measures which are PD, PF, F1, AUC, Accuracy, G-measure, and MCC successively. The 1-st element corresponds to the results of deep representations, and the 2-nd one corresponds to the results of shallow metrics.
4. Models\_mean\_std\_p\_h\_delta\_DeepVSshollow.xls An excel file including five sheets, one sheet includes a 1\*7 array. 1~5 sheets correspond to the performance (meanstd) of deep representations, the performance (meanstd) of shallow metrics, p-value between deep representations and shallow metrics, h-value, and cliff’s delta, respectively.

# RQ2: Proposed TSE VS Baselines

1. Acronyms and Abbreviations

AB AdaBoost

B Bagging

RF Random Forest

TSE The proposed two-stage ensemble learning approach

1. File Content Description
2. run\_i.mat (i=1,2,..,50) A standard data format file for MATLAB, it’s a cell matrix which has five elements. Each element is a 2\*2 confusion matrix, [TN,FP; FN,TP]. i-th confusion matrix corresponds to the results of i-th fold run (5-fold CV is used in our paper). The average result of 5 folds CV is regarded as the result of one run.
3. ‘dataset name’.mat It’s a cell array including four cell elements, each element is a 50\*7 array where 50 means the total runs, and 7 denotes the performance measures which are PD, PF, F1, AUC, Accuracy, G-measure, and MCC successively. 1~4 elements correspond to the performance of TSE, RF, B, and AB, respectively.
4. ‘dataset name’\_mean\_std\_p\_h\_delta\_cv5\_unsolid\_r2modify04\_1runCV.xls An excel file including four sheets. 1~5 sheets correspond to the performance (meanstd) of four models (i.e., TSE, RF, B, and AB), p-value, h-value, and cliff’s delta, respectively. Sheet1 includes a 4\*7 array, sheet2 includes a 3\*7 array in which i-th row denotes the results of p-value between TSE and i-th baseline model, sheet3 includes a 3\*7 array, and sheet4 also includes a 3\*7 array.

# RQ3: Proposed SDAEsTSE VS Baselines

1. Acronyms and Abbreviations

AB AdaBoost

BN Bayesian Network

NB Naive Bayes

RF Random Forest

ST The proposed SDAEsTSE

1. File Content Description
2. i.mat (i=1,2,..,50) A standard data format file for MATLAB, it’s a cell matrix which has five elements. Each element is a 2\*2 confusion matrix, [TN,FP; FN,TP]. i-th confusion matrix corresponds to the results of i-th fold run (5-fold CV is used in our paper). The average result of 5 folds CV is regarded as the result of one run.
3. Models.mat It’s a cell array including five cell elements, each element is a 50\*7 array where 50 means the total runs, and 7 denotes the performance measures which are PD, PF, F1, AUC, Accuracy, G-measure, and MCC successively. 1~5 elements correspond to the performance of ST, BN, RF, AB, and NB, respectively.
4. Models\_mean\_std\_p\_h\_delta\_cv5\_unsolid\_r2modify04\_1runCV.xls An excel file including four sheets. 1~5 sheets correspond to the five models (i.e., ST, BN, RF, AB, and NB) performance (meanstd), p-value, h-value, and cliff’s delta, respectively. Sheet1 is a 5\*7 array, sheet2 is a 4\*7 array in which i-th row denotes the results of p-value between ST and i-th baseline model, sheet3 includes a 4\*7 array, and sheet4 also includes a 4\*7 array.

# Use R or Python to read i.mat file

Each standard MATLAB data file (i.e., i.mat, i=1,2,…,50) includes 5 confusion matrices corresponding to 5 folds cross-validation, and the .mat file also can be easily rad by Python or R as follows:

1. **Python**

import scipy.io as scio

dataFile = 'D://1.mat'

data = scio.loadmat(dataFile)

cm = data['run\_1']

temp = cm[0]

cm1 = temp[0] # the first confusion matrix [TN,FP;FN,TP]

cm2 = temp[1] # the second confusion matrix [TN,FP;FN,TP]

cm3 = temp[2]

cm4 = temp[3]

cm5 = temp[4]

tn = float(cm1[0,0])

fp = float(cm1[0,1])

fn = float(cm1[1,0])

tp = float(cm1[1,1])

precision = tp/(tp+fp)

pd = tp/(tp+fn) # PD also be called as recall

f1 = (2\*precision\*pd)/(precision+pd) # F1 of the first confusion matrix

1. **R**

>install.packages("R.matlab")

>library("R.matlab")

>pathname<-file.path('D:/1.mat')

>data<-readMat(pathname)

>cm1=data[[1]][1][[1]][[1]]; cm2=data[[1]][2][[1]][[1]]; cm3=data[[1]][3][[1]][[1]];

>cm4=data[[1]][4][[1]][[1]]; cm5=data[[1]][5][[1]][[1]]

>tn=cm1[1,1];fp=cm1[1,2];fn=cm1[2,1];tp=cm1[2,2];precision=tp/(tp+fp);pd=tp/(tp+fn);

>f1=(2\*precision\*pd)/(precision+pd)

>print(f1)