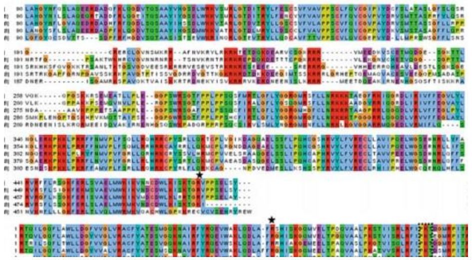
BRNO UNIVERSITY OF TECHNOLOGY

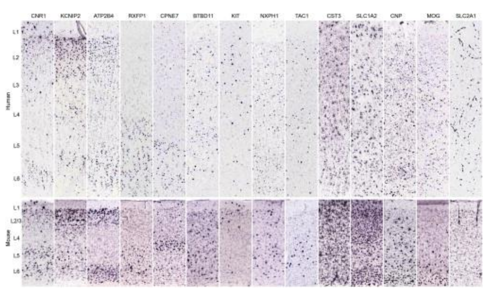
Some aspects of decision support systems: application to differential diagnosis of Parkinson’s Disease and **feature selection**

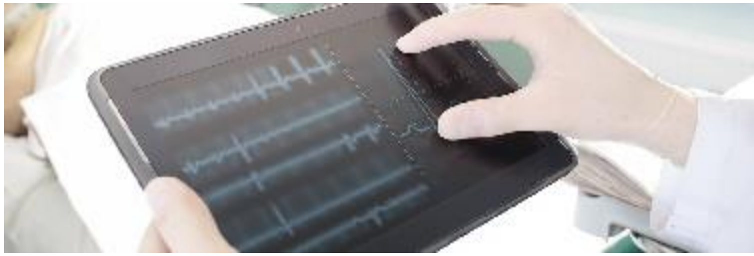
Peter Drotár

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Support of Interdisciplinary Excellence Research Teams Establisment at BUT. CZ.1.07/2.3.00/30.0005

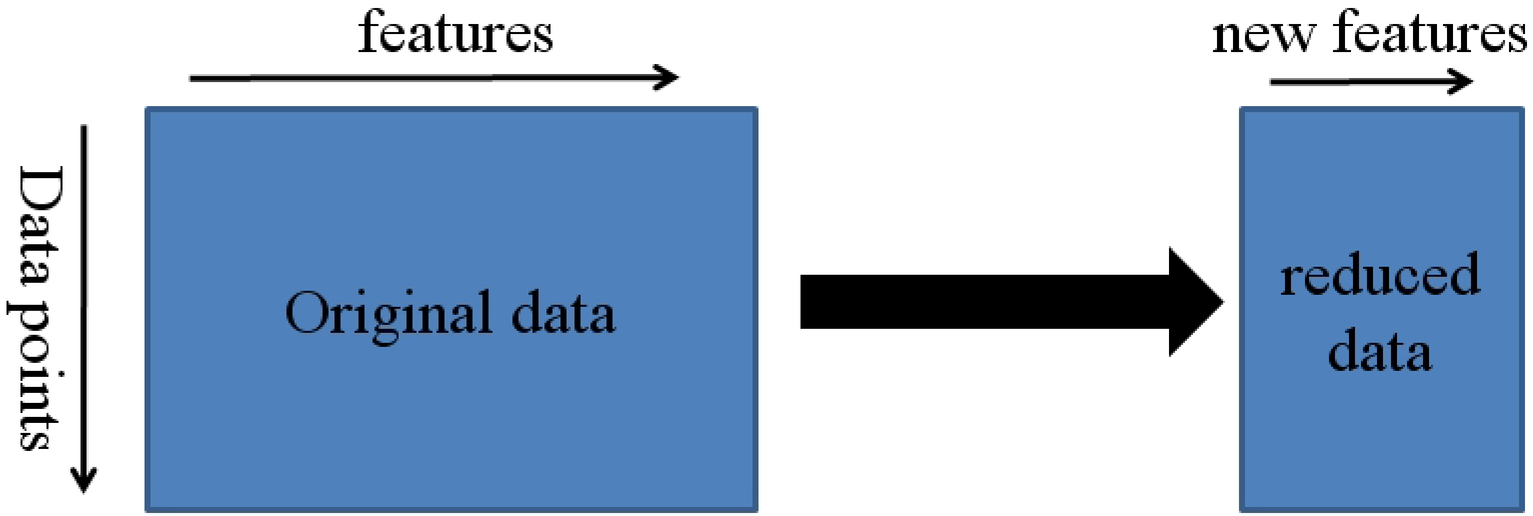






High dimensional data

- Feature Selection techniques select a subset of features from the input which can efficiently describe the input data while reducing effects from noise or irrelevant features and still provide good prediction results.



- problem of feature selection stability
- FS stability - *the robustness of the feature preferences of FS algorithm to differences in training sets drawn from the same generating distribution* [Kalousis et al, 2007]

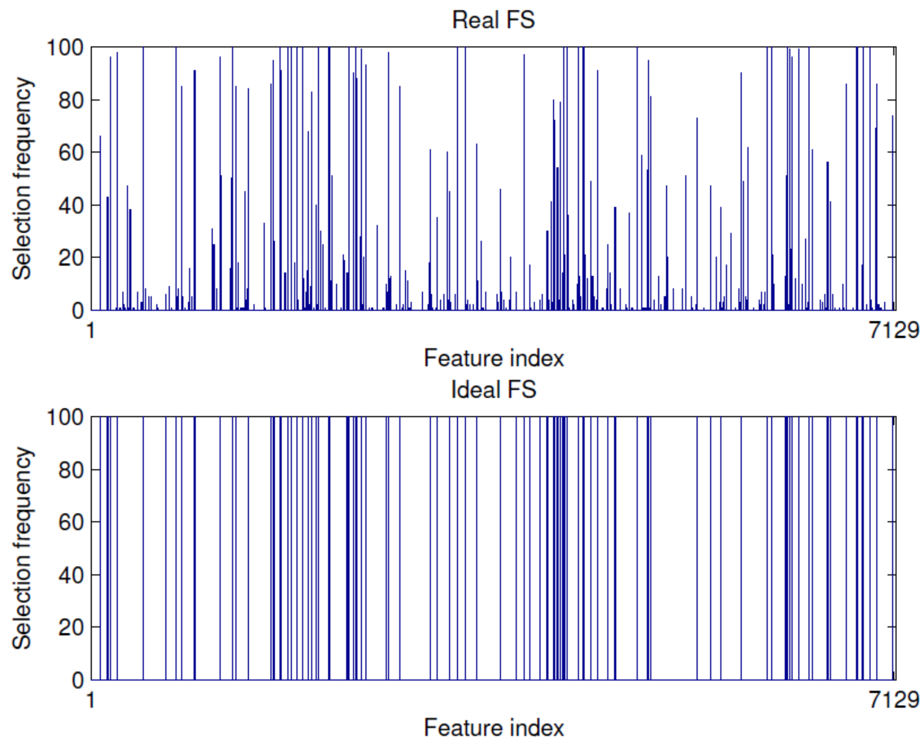


Figure: 100 runs of FS algorithm. FS selects 100 out of 7129 features in each run.

Feature selection and stability of FS – simple example

Assume database of N samples and 10 features (f1, f2, f3, f4, f5, f6, f7, f8, f9, f10).
GOAL: Select 5 most significant features.

Run FS algorithm X with output : f1 f3 f4 f7 f9
Run FS algorithm X with output : f2 f3 f4 f6 f8
Run FS algorithm X with output : f1 f2 f4 f8 f9
Run FS algorithm X with output : f1 f3 f4 f7 f8
Run FS algorithm X with output : f2 f6 f7 f8 f9

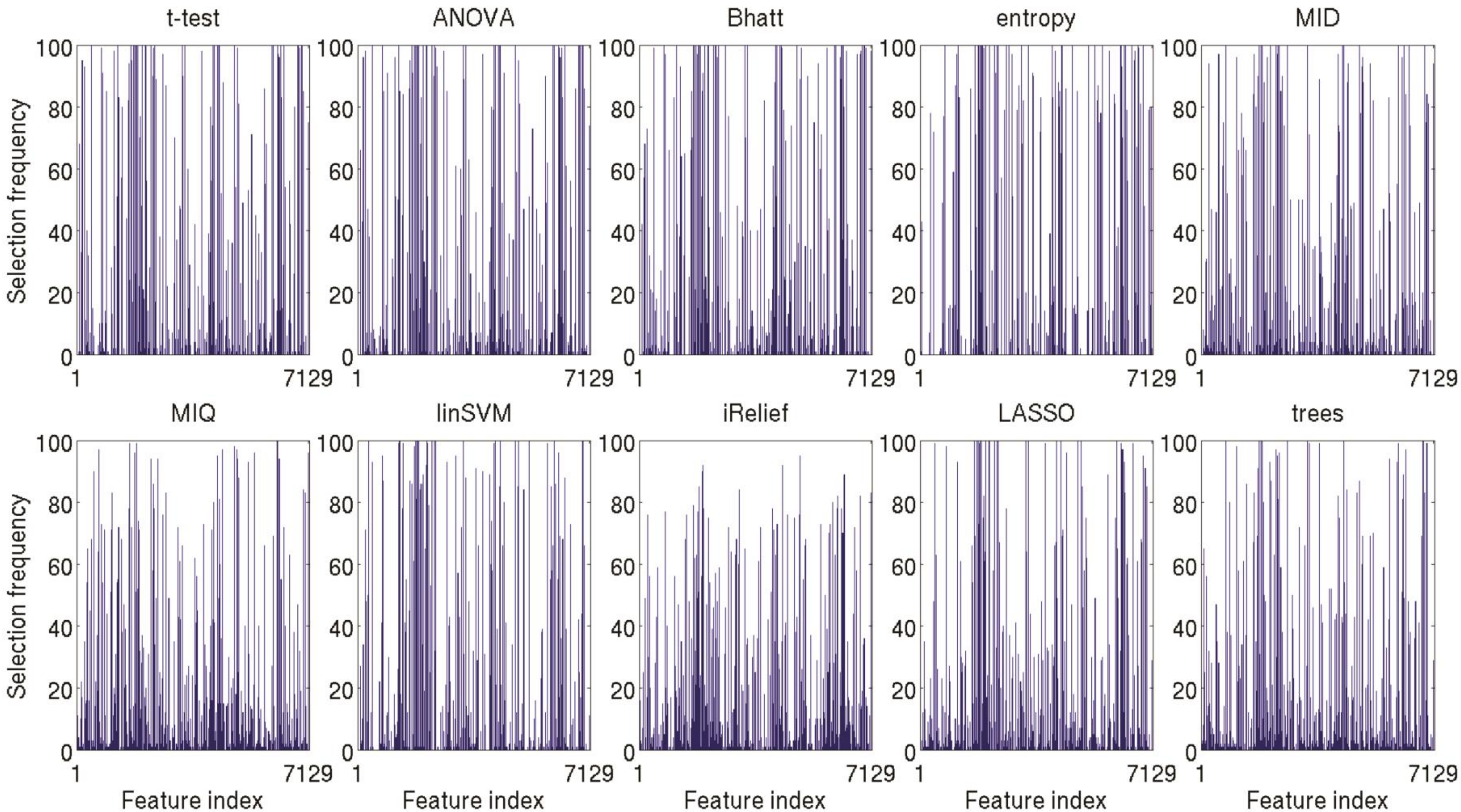
Which features are really significant and how they influence prediction?

Feature selection techniques:

- t-test FS (univariate)
- ANOVA (univariate)
- Bhattacharyya distance (univariate)
- entropy (univariate)
- MRMR – MID
- MRMR – MIQ
- linear SVM
- iterative Relief
- LASSO
- tree

Biomedical datasets:

Dataset name	source	# samples	# features
B2006	Burczynski [1]	127	22,283
C2006	Chowdary [2]	104	22,283
G1999	Golub [3]	72	7129
G2002	Gordon [4]	181	12,533
D2013	Drotar [5]	75	204
T2014	Tsanas [6]	126	309


t-test
ANOVA
Bhatt
entropy
MID
MIQ
linSVM
iRelief
LASSO
trees
Selection frequency
0
20
40
60
80
100
1
7129
Feature index

- Stability measure:
  - Kuncheva index [Kuncheva, 2007]
  - Weighted Consistency index [Somol et al, 2010]

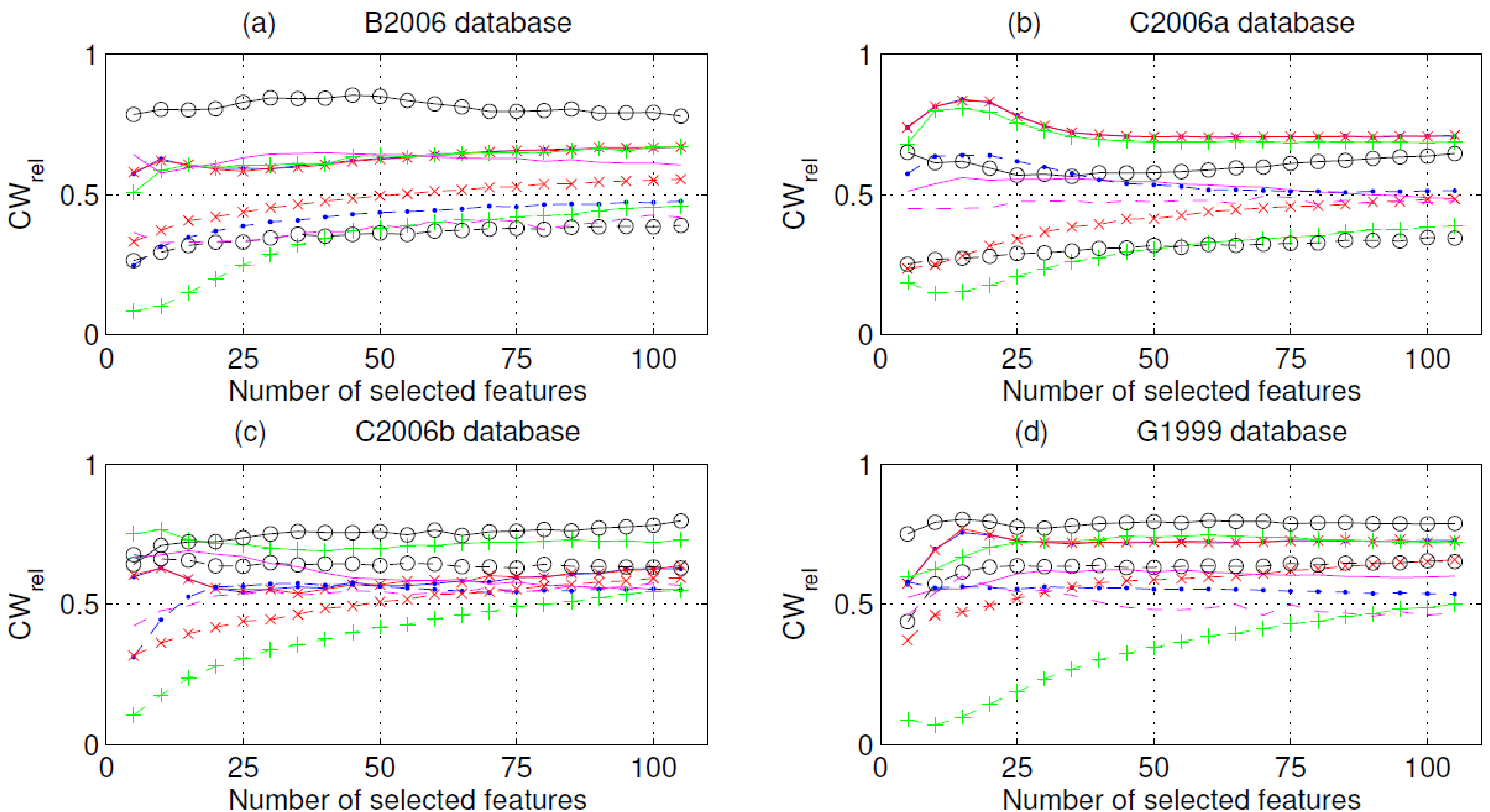
Table 2: Stability for different FS methods measured by Kuncheva index κ

dtb	t-test	ANOVA	Bhatt	entropy	MID	MIQ	linSVM	iRelief	LASSO	Tree
B2006	0.67	0.68	0.66	**0.79**	0.54	0.45	0.39	0.40	0.61	0.47
C2006a	**0.71**	**0.71**	0.68	0.65	0.48	0.37	0.34	0.46	0.48	0.51
C2006b	0.63	0.63	0.73	**0.79**	0.58	0.54	0.62	0.56	0.55	0.54
G1999	0.71	0.71	0.71	**0.78**	0.65	0.48	0.65	0.45	0.59	0.54
G2002	0.77	0.78	0.78	**0.83**	0.73	0.50	0.46	0.65	0.66	0.59
T2003	0.45	0.45	0.46	**0.68**	0.35	0.29	0.13	0.46	0.41	0.25
D2013	0.43	0.42	0.40	0.45	0.54	0.46	0.50	0.37	**0.55**	0.36
T2014	0.70	0.72	0.67	**0.90**	0.66	0.64	0.49	0.53	0.80	0.57
average	0.63	0.64	0.64	**0.74**	0.57	0.47	0.45	0.48	0.58	0.48

Table 3: Stability for different FS methods measured by weighted consistency *CW*

dtb	t-test	ANOVA	Bhatt	entropy	MID	MIQ	linSVM	iRelief	LASSO	Tree
B2006	0.67	0.68	0.66	**0.79**	0.55	0.45	0.39	0.40	0.61	0.47
C2006a	0.71	**0.72**	0.68	0.65	0.48	0.37	0.34	0.46	0.48	0.51
C2006b	0.63	0.63	0.73	**0.79**	0.58	0.55	0.62	0.56	0.55	0.55
G1999	0.71	0.72	0.72	**0.79**	0.65	0.49	0.65	0.46	0.59	0.54
G2002	0.77	0.78	0.78	**0.84**	0.73	0.50	0.47	0.65	0.66	0.59
T2003	0.45	0.46	0.47	**0.68**	0.36	0.29	0.14	0.46	0.41	0.25
D2013	0.71	0.70	0.69	0.72	**0.77**	0.72	0.74	0.68	**0.77**	0.67
T2014	0.79	0.81	0.78	**0.94**	0.77	0.75	0.65	0.68	0.87	0.71
average	0.68	0.69	0.69	**0.77**	0.61	0.51	0.50	0.54	0.61	0.54

- FS stability as a function of number of selected features
- Stability measure : relative weighted consistency index



P. Drotár (Brno University of Technology)

- FS **similarity**
- Similarity measure : intersystem Kuncheva index

Table 4: Similarity of FS techniques expressed by intersystem Kunacheva index I_K. G1999 database.

FS	t-test	ANOVA	Bhatt	entropy	MID	MIQ	linSVM	iRelief	LASSO	Tree
t-test		0.96	0.87	0.16	0.59	0.42	0.52	0.20	0.60	0.59
ANOVA	0.96		0.88	0.16	0.58	0.42	0.51	0.19	0.59	0.59
Bhatt	0.87	0.88		0.20	0.62	0.44	0.44	0.18	0.57	0.60
entropy	0.16	0.16	0.20		0.26	0.23	0.08	0.04	0.14	0.16
MID	0.59	0.58	0.62	0.26		0.71	0.35	0.20	0.44	0.49
MIQ	0.42	0.42	0.44	0.23	0.71		0.32	0.17	0.34	0.38
linSVM	0.52	0.51	0.44	0.08	0.35	0.32		0.36	0.42	0.36
iRelief	0.20	0.19	0.18	0.04	0.20	0.17	0.36		0.27	0.21
LASSO	0.60	0.59	0.57	0.14	0.44	0.34	0.42	0.27		0.51
Tree	0.59	0.59	0.60	0.16	0.49	0.38	0.36	0.21	0.51	

- FS influence on **prediction accuracy**
- Accuracy measure : Matthews correlation coefficient

Table 6: MCC performance

dtb	classifier	t-test	ANOVA	Bhatt	entropy	MID	MIQ	linSVM	iRelief	LASSO	Tree
B2006	Ada	91.6	91.4	93.4	90.9	**94.9**	94.4	93.7	89.4	93.6	93.4
C2006a	Ada	78.0	78.5	79.5	78.1	**93.8**	81.3	78.1	67.8	76.5	80.7
C2006b	Ada	**96.4**	80.6	**96.4**	94.4	94.2	**96.4**	92.1	70.1	78.8	78.4
G1999	Ada	92.9	94.8	**97.5**	97.3	94.5	97.3	93.8	96.4	96.5	96.4
G2002	Ada	97.7	97.5	97.7	95.9	97.7	**100.0**	**100.0**	92.2	96.7	96.4
T2003	Ada	35.4	97.7	32.2	44.8	**98.2**	35.3	31.7	23.6	25.5	27.1
D2013	Ada	19.9	36.9	27.1	49.4	56.6	**61.1**	6.4	55.4	48.8	56.3
T2014	Ada	**78.4**	70.0	73.6	69.4	70.9	72.1	50.5	69.7	70.9	69.5

- entropy based FS appears to be the most stable FS
- Features selected by mRMR techniques helps to achieve highest prediction accuracy
  - however accuracies are comparable