

# AlvsPK Challenge: FACT SHEET

**Title:** Hybrid approach for learning

**Name, address, email:** Mehreen Saeed, FAST National University of Computer & Emerging Science, Lahore, Pakistan, mehreen.saeed@nu.edu.pk

**Acronym of your best entry:** SubmitA

**Reference:**

N/A

**Method:**

- Preprocessing : Standardize, shift-n-scale, normalize
- Feature selection : Removed the sparse features whose percentage of non-zero entries was below as certain number.
- Classification
  - Used mixture models for finding clusters within data. I wrote down my own routine for generating the mixture parameters
  - For binary data, Bernoulli mixture models were used. For continuous data Gaussian mixture models were used
  - Neural network and gentleboost from CLOP / spider was used for learning the conditional class probabilities as a function of class label
  - Did you use ensemble methods? No
  - Did you use “transduction” or learning from the unlabeled test set? No
- Model selection/hyperparameter selection : 5-fold cross-validation was used

**Results:**

Table 1: Our methods best results

Dataset	Entry name	Entry ID	Test BER	Test AUC	Score	Track
ADA	Submit D final	1037	0.181	0.8185	0.1022	Agnos
GINA	Submit A final	1034	0.0495	0.9507	0.3162	Agnos
HIVA	Submit D final	1037	0.305	0.6976	0.4418	Agnos
NOVA	Submit E final	1038	0.0456	0.9552	0.0385	Agnos
SYLVA	Submit C final	1036	0.0094	0.9906	0.2864	Agnos
Overall	Submit D final	1037	0.1194	0.8812	0.2786	Agnos

Table 2: Winning entries of the AlvsPK challenge

Best results agnostic learning track						
Dataset	Entrant name	Entry name	Entry ID	Test BER	Test AUC	Score
ADA	Roman Lutz	LogitBoost with trees	13, 18	0.166	0.9168	0.002
GINA	Roman Lutz	LogitBoost/Doubleboost	892, 893	0.0339	0.9668	0.2308
HIVA	Vojtech Franc	RBF SVM	734, 933, 934	0.2827	0.7707	0.0763
NOVA	Mehreen Saeed	Submit E final	1038	0.0456	0.9552	0.0385
SYLVA	Roman Lutz	LogitBoost with trees	892	0.0062	0.9938	0.0302
<b>Overall</b>	Roman Lutz	LogitBoost with trees	892	0.1117	0.8892	0.1431
Best results prior knowledge track						
Dataset	Entrant name	Entry name	Entry ID	Test BER	Test AUC	Score
ADA	Marc Boulle	Data Grid	920, 921, 1047	0.1756	0.8464	0.0245
GINA	Vladimir Nikulin	vn2	1023	0.0226	0.9777	0.0385
HIVA	Chloe Azencott	SVM	992	0.2693	0.7643	0.008
NOVA	Jorge Sueiras	Boost mix	915	0.0659	0.9712	0.3974
SYLVA	Roman Lutz	Doubleboost	893	0.0043	0.9957	0.005
<b>Overall</b>	Vladimir Nikulin	vn3	1024	0.1095	0.8949	0.095967

- quantitative advantages : Dimensionality reduction
- qualitative advantages : Novel approach that combines both the advantages of a generative and a discriminative classifier.

**Code:** If CLOP or the Spider were used, fill out the table:

Dataset	Spider command used to build the model
ADA	
GINA	
HIVA	
NOVA	
SYLVA	

I used spider/clop commands and added my own objects to the spider library. I implemented my own version of the expectation maximization algorithm in C++ and called this routine from matlab.

**Keywords:** Put at *least one keyword in each category*. Try some of the following keywords and add your own:

- Preprocessing or feature construction: centering, scaling, standardization
- Feature selection approach: frequency count
- Feature selection engine: Very simple matlab routine
- Feature selection search: Brute force
- Feature selection criterion: data statistics
- Classifier: Neural networks, SVM, boosting, mixture model
- Hyper-parameter selection: cross-validation
- Other: post-processing, 'bias' option