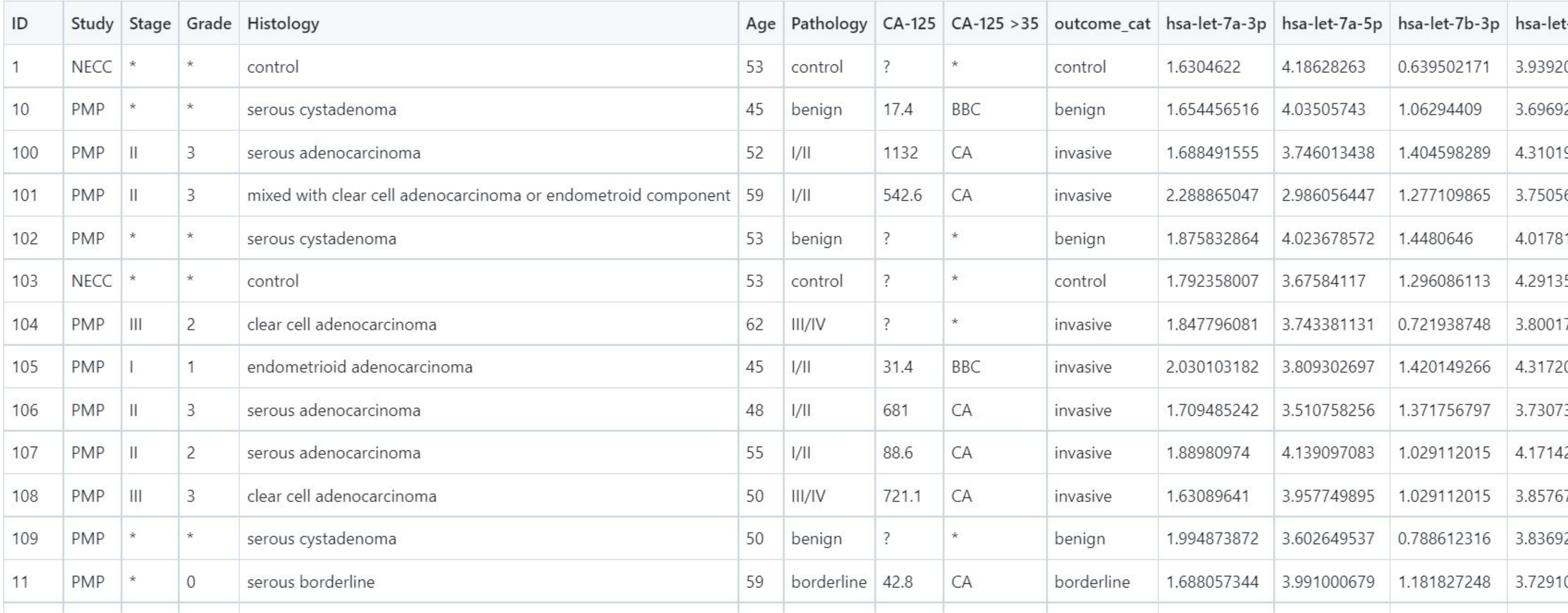
Paper3: Diagnostic potential for a serum miRNA neural network for detection of ovarian cancer

Team member:

章峻福 110753503
葉冠宏 108753208

Inputs

Our inout data:



ID	Study	Stage	Grade	Histology	Age	Pathology	CA-125	CA-125 >35	outcome_cat	hsa-let-7a-3p	hsa-let-7a-5p	hsa-let-7b-3p	hsa-let-
1	NECC	*	*	control	53	control	?	*	control	1.6304622	4.18628263	0.639502171	3.9392
10	PMP	*	*	serous cystadenoma	45	benign	17.4	BBC	benign	1.654456516	4.03505743	1.06294409	3.6969
100	PMP	II	3	serous adenocarcinoma	52	I/II	1132	CA	invasive	1.688491555	3.746013438	1.404598289	4.3101
101	PMP	II	3	mixed with clear cell adenocarcinoma or endometroid component	59	I/II	542.6	CA	invasive	2.288865047	2.986056447	1.277109865	3.7505
102	PMP	*	*	serous cystadenoma	53	benign	?	*	benign	1.875832864	4.023678572	1.4480646	4.0178
103	NECC	*	*	control	53	control	?	*	control	1.792358007	3.67584117	1.296086113	4.2913
104	PMP	III	2	clear cell adenocarcinoma	62	III/IV	?	*	invasive	1.847796081	3.743381131	0.721938748	3.8001
105	PMP	I	1	endometrioid adenocarcinoma	45	I/II	31.4	BBC	invasive	2.030103182	3.809302697	1.420149266	4.3172
106	PMP	II	3	serous adenocarcinoma	48	I/II	681	CA	invasive	1.709485242	3.510758256	1.371756797	3.7307
107	PMP	II	2	serous adenocarcinoma	55	I/II	88.6	CA	invasive	1.88980974	4.139097083	1.029112015	4.1714
108	PMP	III	3	clear cell adenocarcinoma	50	III/IV	721.1	CA	invasive	1.63089641	3.957749895	1.029112015	3.8576
109	PMP	*	*	serous cystadenoma	50	benign	?	*	benign	1.994873872	3.602649537	0.788612316	3.8369
11	PMP	*	0	serous borderline	59	borderline	42.8	CA	borderline	1.688057344	3.991000679	1.181827248	3.7291

Note: We didn’t do further data preprocessing since the author of the paper has done it for us.

Feature

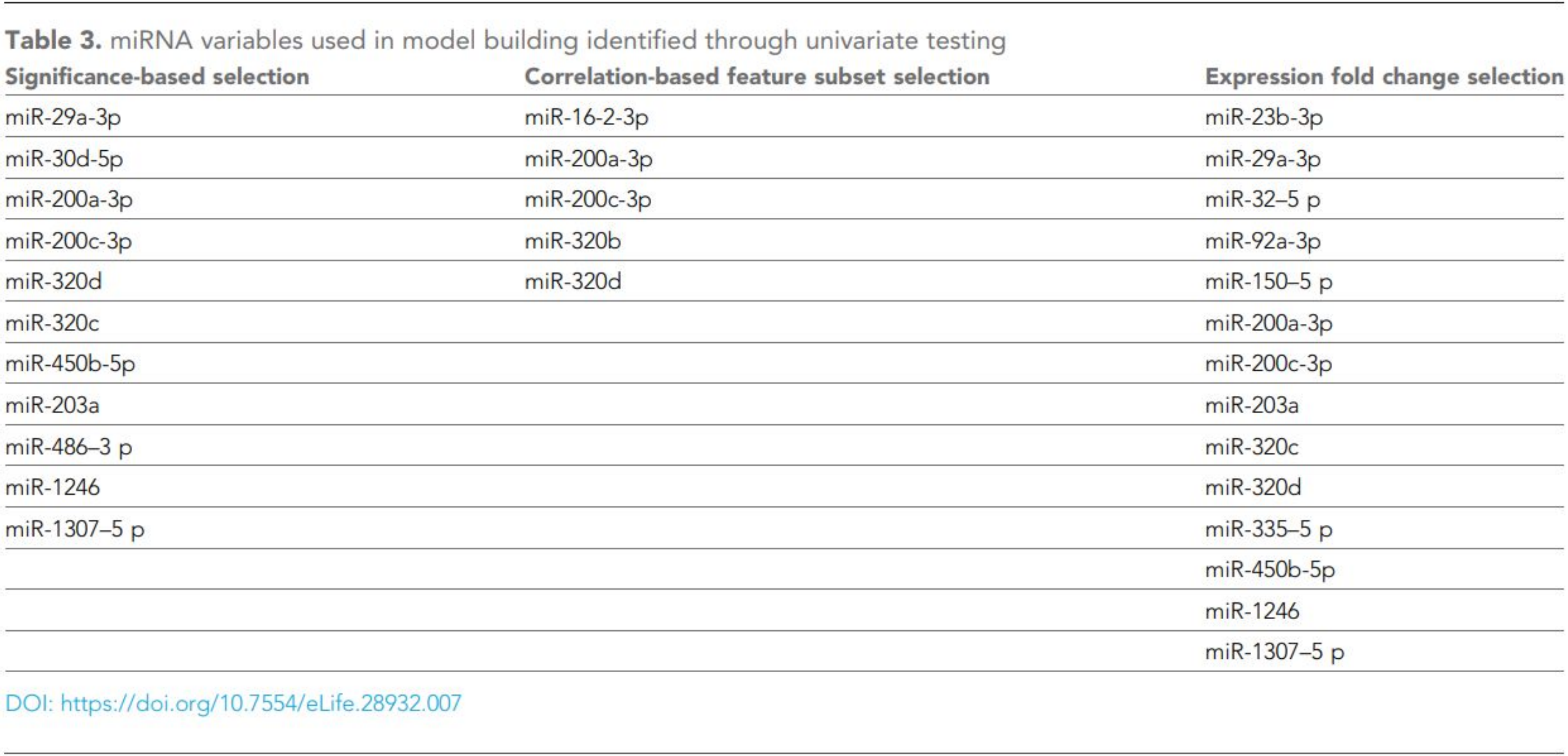


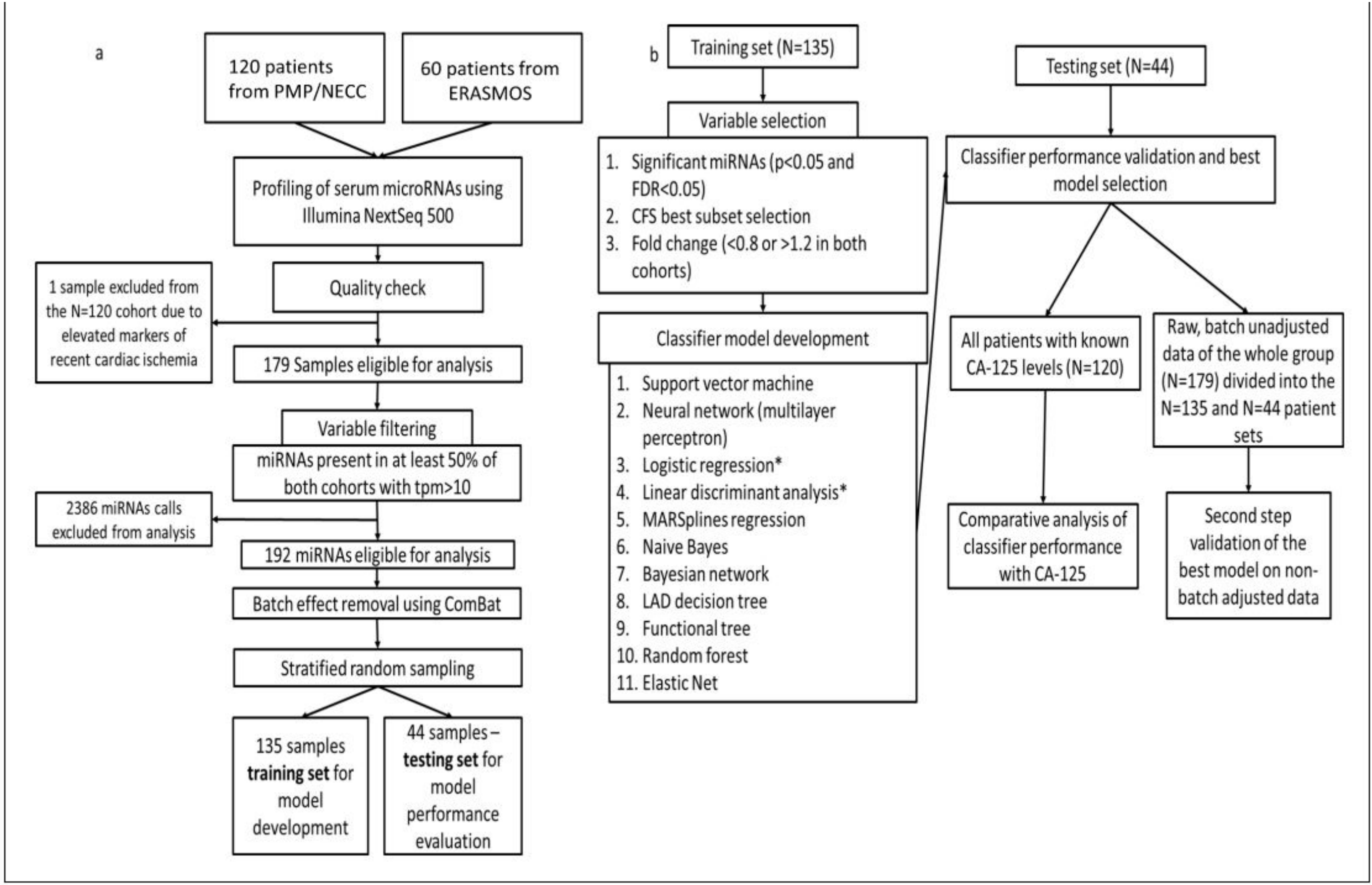
Table 3. miRNA variables used in model building identified through univariate testing

Significance-based selection	Correlation-based feature subset selection	Expression fold change selection
miR-29a-3p	miR-16-2-3p	miR-23b-3p
miR-30d-5p	miR-200a-3p	miR-29a-3p
miR-200a-3p	miR-200c-3p	miR-32–5 p
miR-200c-3p	miR-320b	miR-92a-3p
miR-320d	miR-320d	miR-150–5 p
miR-320c		miR-200a-3p
miR-450b-5p		miR-200c-3p
miR-203a		miR-203a
miR-486–3 p		miR-320c
miR-1246		miR-320d
miR-1307–5 p		miR-335–5 p
		miR-450b-5p
		miR-1246
		miR-1307–5 p

DOI: https://doi.org/10.7554/eLife.28932.007

Workflow

Our flowchart:



a

120 patients from PMP/NECC

60 patients from ERASMOS

Profiling of serum microRNAs using Illumina NextSeq 500

Quality check

1 sample excluded from the N=120 cohort due to elevated markers of recent cardiac ischemia

179 Samples eligible for analysis

Variable filtering

miRNAs present in at least 50% of both cohorts with tpm>10

2386 miRNAs calls excluded from analysis

192 miRNAs eligible for analysis

Batch effect removal using ComBat

Stratified random sampling

135 samples **training set** for model development

44 samples – **testing set** for model performance evaluation

b

Training set (N=135)

Variable selection

1. Significant miRNAs (p<0.05 and FDR<0.05)
2. CFS best subset selection
3. Fold change (<0.8 or >1.2 in both cohorts)

Classifier model development

1. Support vector machine
2. Neural network (multilayer perceptron)
3. Logistic regression*
4. Linear discriminant analysis*
5. MARSplines regression
6. Naive Bayes
7. Bayesian network
8. LAD decision tree
9. Functional tree
10. Random forest
11. Elastic Net

Testing set (N=44)

Classifier performance validation and best model selection

All patients with known CA-125 levels (N=120)

Comparative analysis of classifier performance with CA-125

Raw, batch unadjusted data of the whole group (N=179) divided into the N=135 and N=44 patient sets

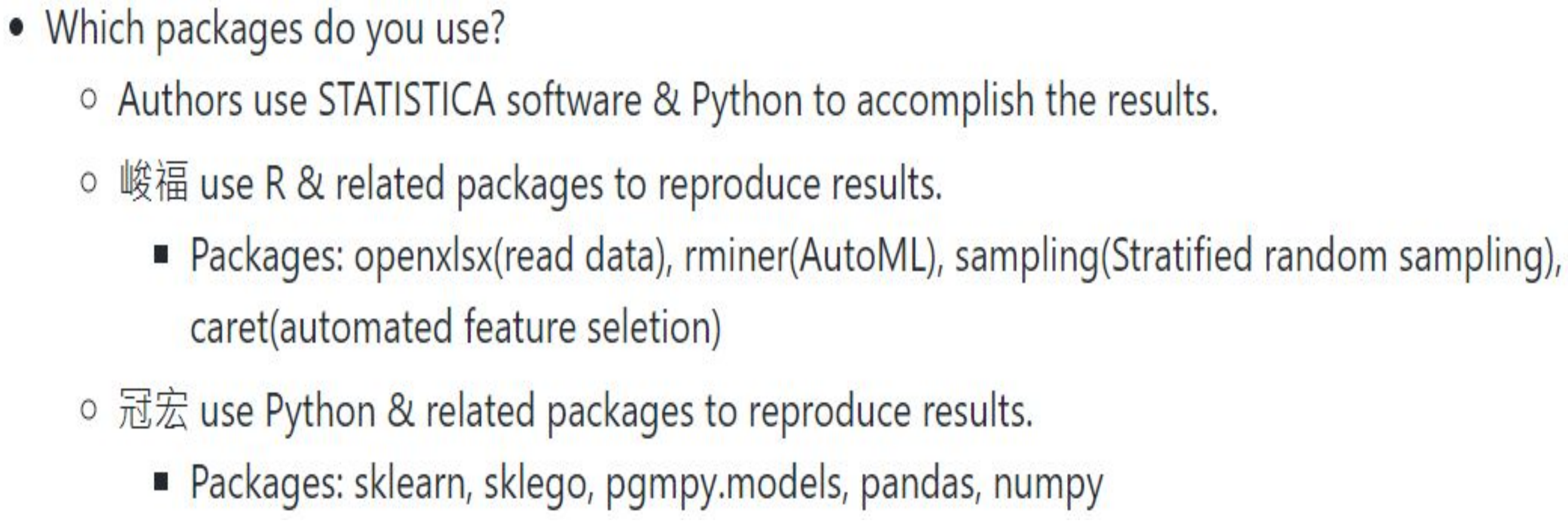
Second step validation of the best model on non-batch adjusted data

Figure 1. Flowchart of study design. (**a**) Protocol for miRNA sequencing, filtering, batch adjustment and separation into the training and testing sets. (**b**) Protocol for model development and testing.

DOI: https://doi.org/10.7554/eLife.28932.003

Packages

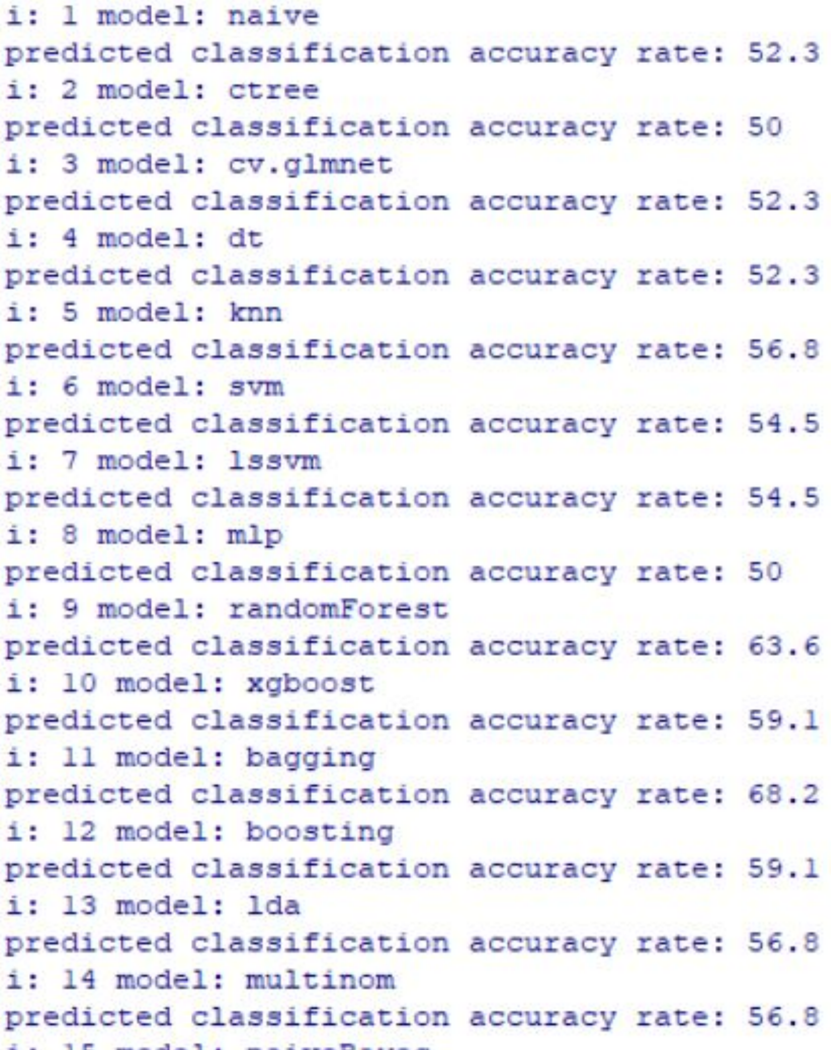
Which tools do you use?(related publication):



- Which packages do you use?
  - Authors use STATISTICA software & Python to accomplish the results.
  - 峻福 use R & related packages to reproduce results.
    - Packages: openxlsx(read data), rminer(AutoML), sampling(Stratified random sampling), caret(automated feature seletion)
  - 冠宏 use Python & related packages to reproduce results.
    - Packages: sklearn, sklego, pgmpy.models, pandas, numpy

Results

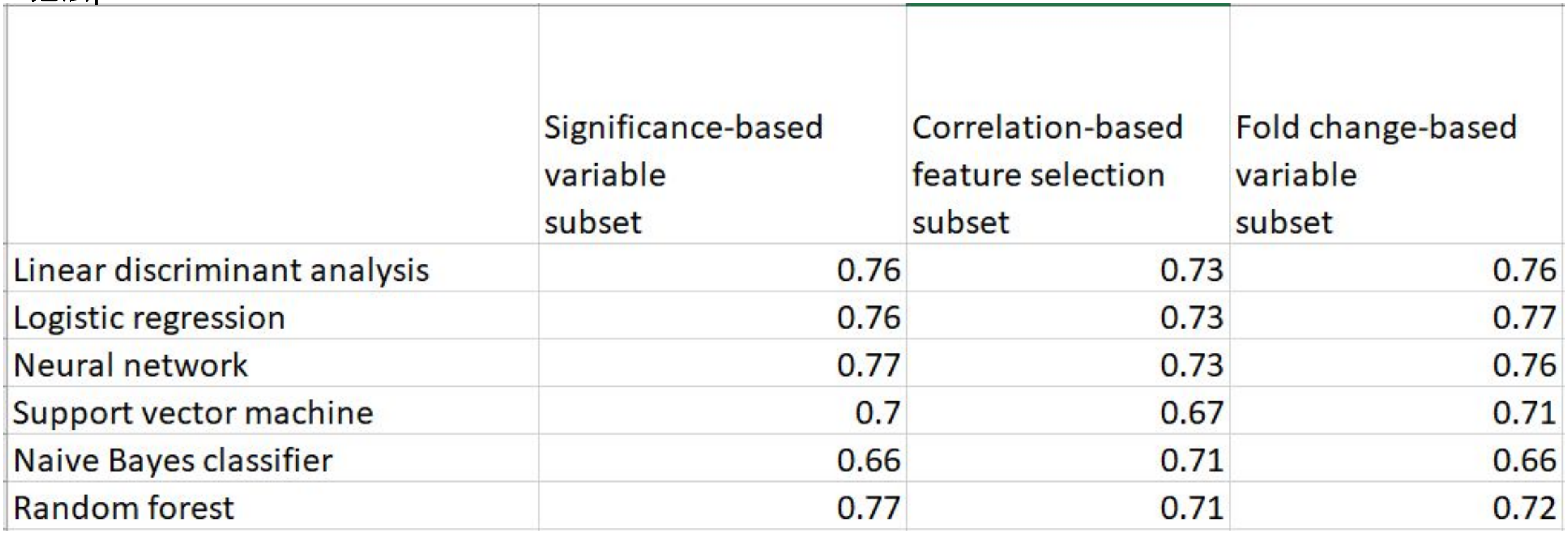
峻福part:



```
i: 1 model: naive
predicted classification accuracy rate: 52.3
i: 2 model: ctree
predicted classification accuracy rate: 50
i: 3 model: cv.glmnet
predicted classification accuracy rate: 52.3
i: 4 model: dt
predicted classification accuracy rate: 52.3
i: 5 model: knn
predicted classification accuracy rate: 56.8
i: 6 model: svm
predicted classification accuracy rate: 54.5
i: 7 model: lssvm
predicted classification accuracy rate: 54.5
i: 8 model: mlp
predicted classification accuracy rate: 50
i: 9 model: randomForest
predicted classification accuracy rate: 63.6
i: 10 model: xgboost
predicted classification accuracy rate: 59.1
i: 11 model: bagging
predicted classification accuracy rate: 68.2
i: 12 model: boosting
predicted classification accuracy rate: 59.1
i: 13 model: lda
predicted classification accuracy rate: 56.8
i: 14 model: multinom
predicted classification accuracy rate: 56.8
i: 15 model: naiveBayes
```

Results

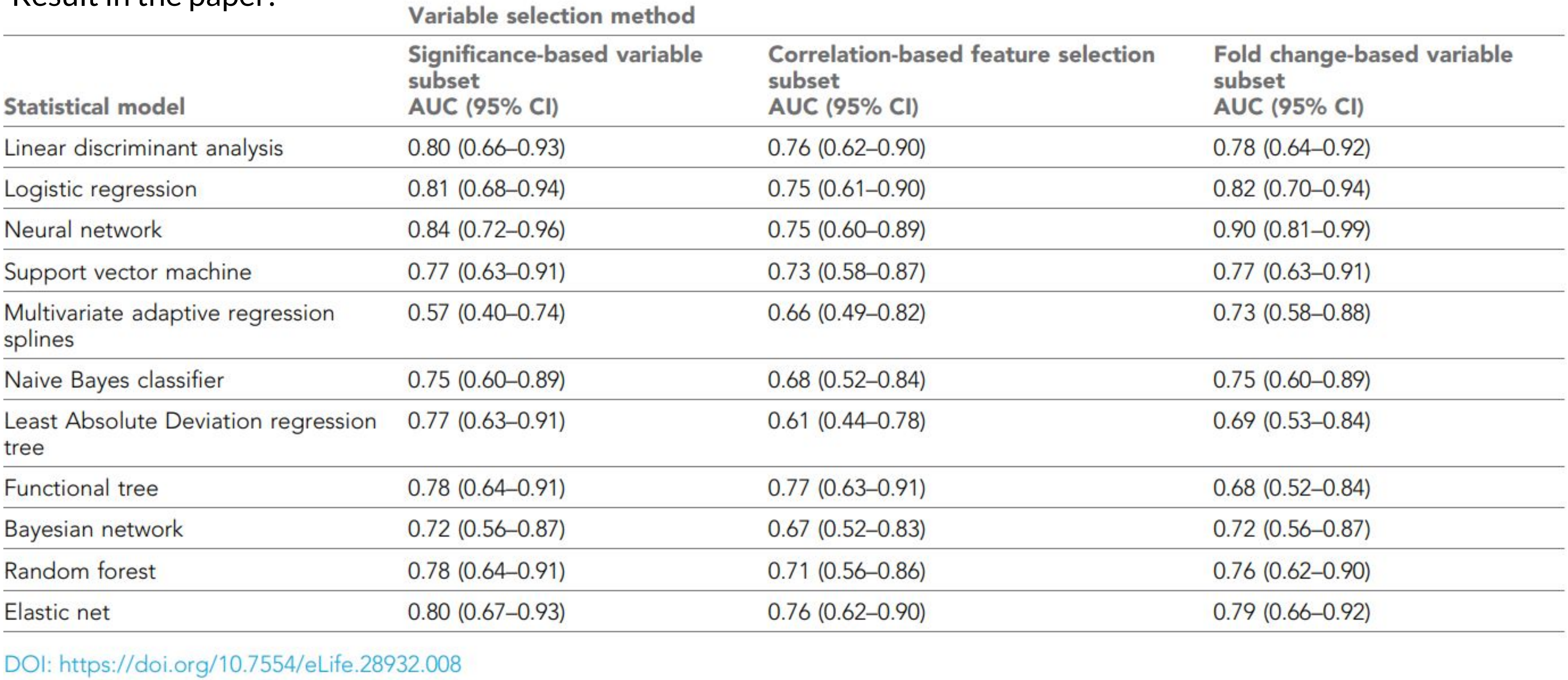
冠宏part:



	Significance-based variable subset	Correlation-based feature selection subset	Fold change-based variable subset
Linear discriminant analysis	0.76	0.73	0.76
Logistic regression	0.76	0.73	0.77
Neural network	0.77	0.73	0.76
Support vector machine	0.7	0.67	0.71
Naive Bayes classifier	0.66	0.71	0.66
Random forest	0.77	0.71	0.72

Results

Result in the paper:



	Variable selection method		
Statistical model	Significance-based variable subset AUC (95% CI)	Correlation-based feature selection subset AUC (95% CI)	Fold change-based variable subset AUC (95% CI)
Linear discriminant analysis	0.80 (0.66–0.93)	0.76 (0.62–0.90)	0.78 (0.64–0.92)
Logistic regression	0.81 (0.68–0.94)	0.75 (0.61–0.90)	0.82 (0.70–0.94)
Neural network	0.84 (0.72–0.96)	0.75 (0.60–0.89)	0.90 (0.81–0.99)
Support vector machine	0.77 (0.63–0.91)	0.73 (0.58–0.87)	0.77 (0.63–0.91)
Multivariate adaptive regression splines	0.57 (0.40–0.74)	0.66 (0.49–0.82)	0.73 (0.58–0.88)
Naive Bayes classifier	0.75 (0.60–0.89)	0.68 (0.52–0.84)	0.75 (0.60–0.89)
Least Absolute Deviation regression tree	0.77 (0.63–0.91)	0.61 (0.44–0.78)	0.69 (0.53–0.84)
Functional tree	0.78 (0.64–0.91)	0.77 (0.63–0.91)	0.68 (0.52–0.84)
Bayesian network	0.72 (0.56–0.87)	0.67 (0.52–0.83)	0.72 (0.56–0.87)
Random forest	0.78 (0.64–0.91)	0.71 (0.56–0.86)	0.76 (0.62–0.90)
Elastic net	0.80 (0.67–0.93)	0.76 (0.62–0.90)	0.79 (0.66–0.92)

DOI: https://doi.org/10.7554/eLife.28932.008

What do we find?

1. The result Kuan-Hung did has inferior result. =>may be due to statistical bias.

On-line demo

Reproducibility

1.How to document our project? How to maintain our code? How to reproduce our result?
冠宏part: use data “final_dset_combine.csv” and run “test.py”
峻福part: run “f_project_tmp.r”

2.Teamwork coordination
峻福:research on feature engineering, run AutoML methods which is not implemented by the author, edit Github
冠宏:do some package research, implement the methods mentioned in the paper, edit Github, edit slides