Final Project

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

Members : 章峻福110753503
葉冠宏 108753208

1 What is this paper talking about

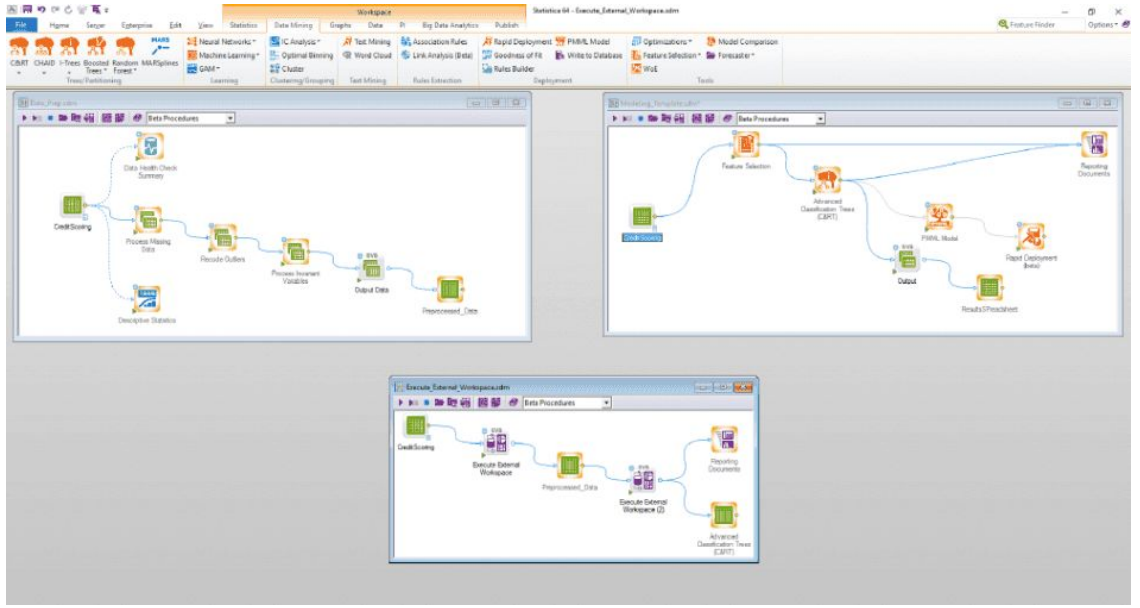
2 The key point what we got

3 Our results

What is this paper talking about

- Multi-class Classification
  - Distinguish 4 class of the ovarian cancer status – control, benign, borderline tumor, invasive cancer
  - Variables: 2,578 microRNAs
- develop a Neural Netwrok to classify patients well
  - Comparision: serum CA125 level
- Traditional ML process
  - Hyperparameter tuning?
  - Feature selection?

STATISTICA



Flowchart of study design

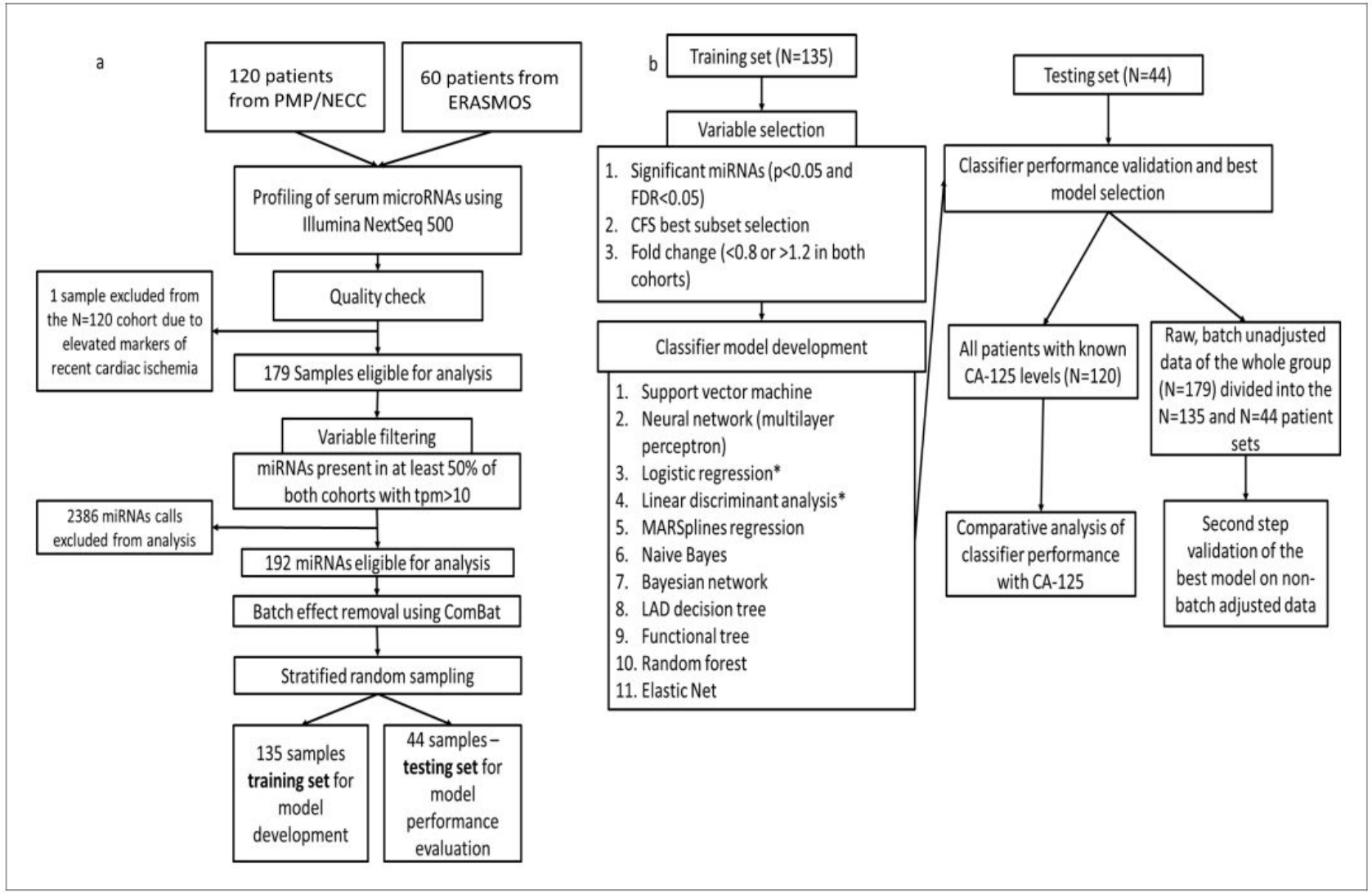
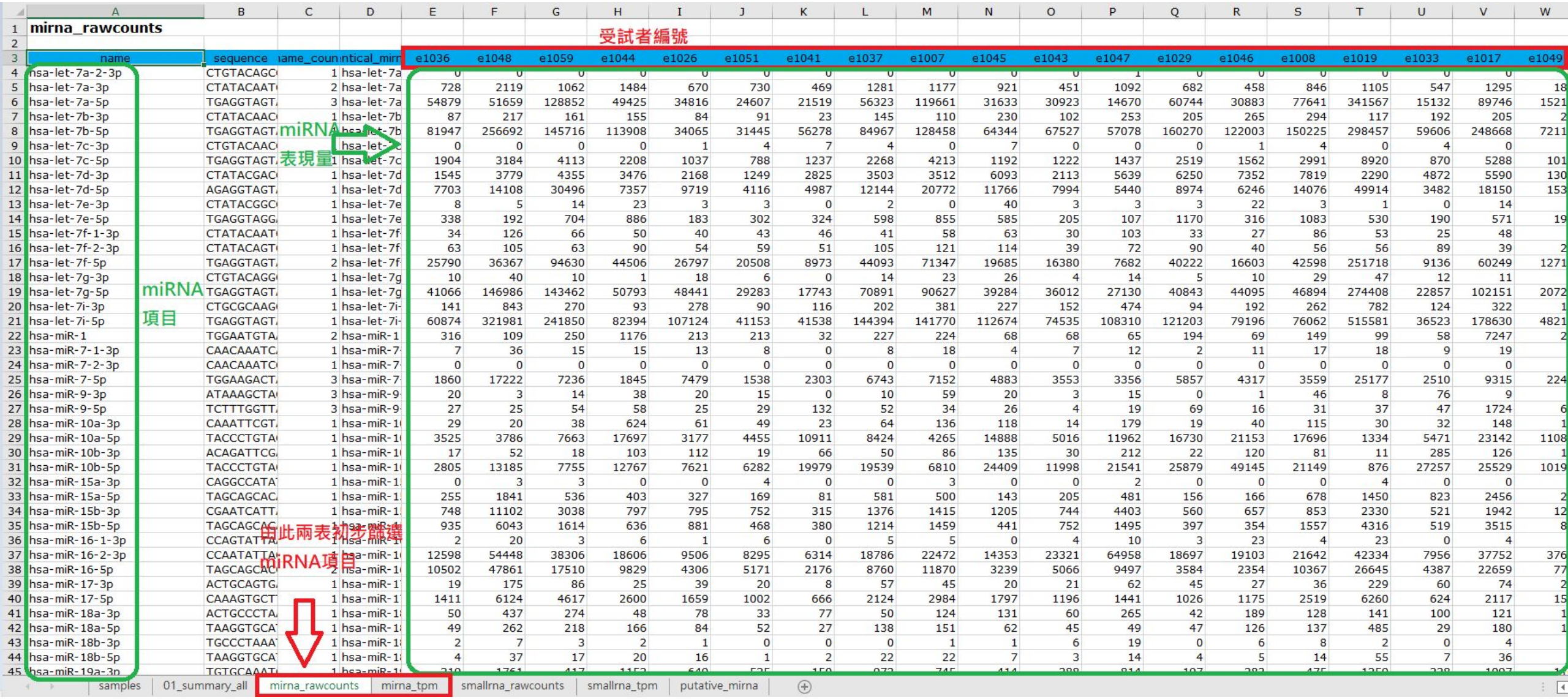


Figure 1. Flowchart of study design. (a) Protocol for miRNA sequencing, filtering, batch adjustment and separation into the training and testing sets. (b) Protocol for model development and testing.

Data Available

- GSE94533
  - Processed RNA-seq data (rawcount & TPM)
- Supplement data
  - elife-28932-supp1-v2.docx
    - clinical data: Age, Study, Cancer Stage, grade, Histology, CA125 levels
  - elife-28932-supp6-v2.xlsx
    - Batch adjusted, log10-transformed miRNA expression data

Progressed data Structure



Data Merged

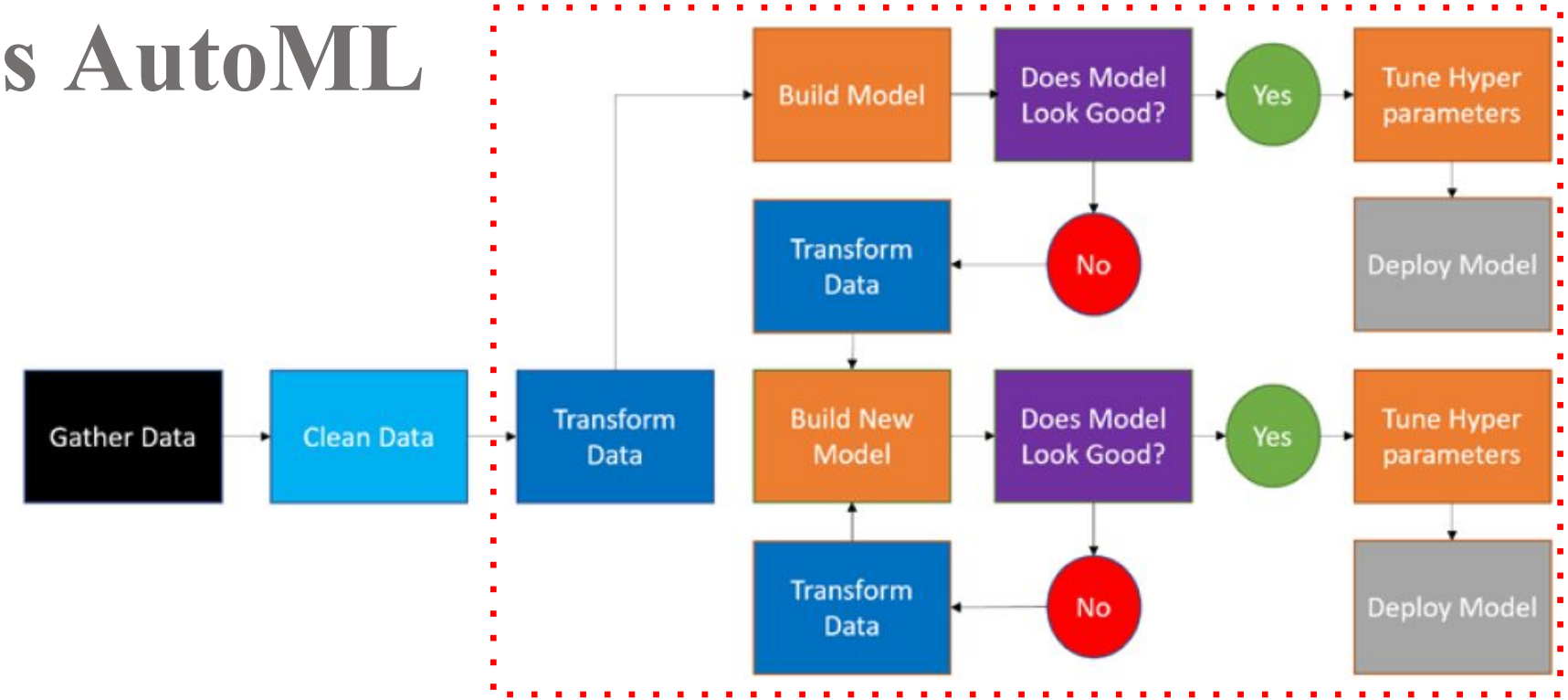
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	ID	Study	Stage	Grade	Histology	Age	Pathology	outcome_c	CA-125	CA-125 >3	hsa-let-7a-	hsa-let-7a-	hsa-let-7b-	hsa-let-7b-	hsa-let-7c-	hsa-let-7d-	hsa-let-7d-	hsa-let-7e-	hsa-let-7f-	hsa-let-7g-	hsa-let-7i-	hsa-miR-1
2	1	NECC	*	*	control	53	control	control	‡	*	1.630462	4.186283	0.639502	3.939203	2.781	2.231032	3.167665	1.804878	4.069749	4.142388	4.29829	1.506029
3	10	PMP	*	*	serous cyst	45	benign	benign	17.4	BBC	1.654457	4.035057	1.062944	3.696924	2.575665	2.693975	3.074113	1.870031	3.986123	3.595796	4.070027	2.155441
4	100	PMP	II	3	serous ader	52	I/II	invasive	1132	CA	1.688492	3.746013	1.404598	4.310191	2.726369	2.833369	2.524331	1.976228	3.641565	3.590177	3.869842	1.226399
5	101	PMP	II	3	mixed with	59	I/II	invasive	542.6	CA	2.288865	2.986056	1.27711	3.750561	2.029923	2.943666	2.322157	1.798989	3.344037	3.35462	3.940302	1.226399
6	102	PMP	*	*	serous cyst	53	benign	benign	‡	*	1.875833	4.023679	1.448065	4.017818	2.864536	3.015019	2.804518	2.457346	3.956777	3.659199	3.916113	1.322702
7	103	NECC	*	*	control	53	control	control	‡	*	1.792358	3.675841	1.296086	4.291353	2.605935	2.679978	2.601779	1.605416	3.520133	3.497688	3.945064	0.577276
8	104	PMP	III	2	clear cell a	62	III/IV	invasive	‡	*	1.847796	3.743381	0.721939	3.800174	2.497354	2.198513	2.611616	2.042312	3.763583	3.384017	3.8527	1.356471
9	105	PMP	I	1	endometric	45	I/II	invasive	31.4	BBC	2.030103	3.809303	1.420149	4.317201	2.703967	2.900169	2.753855	1.989602	3.779855	3.671378	3.98902	1.305231
10	106	PMP	II	3	serous ader	48	I/II	invasive	681	CA	1.709485	3.510758	1.371757	3.730733	2.331247	2.522019	2.409475	1.95892	3.507009	3.230848	3.641276	1.286774
11	107	PMP	II	2	serous ader	55	I/II	invasive	88.6	CA	1.88981	4.139097	1.029112	4.171426	2.847593	2.286256	2.9984	1.890304	4.084689	3.858878	4.219908	1.819692
12	108	PMP	III	3	clear cell a	50	III/IV	invasive	721.1	CA	1.630896	3.95775	1.029112	3.857679	2.58789	2.77785	2.980401	2.435426	3.885882	3.610706	4.055711	1.677874
13	109	PMP	*	*	serous cyst	50	benign	benign	‡	*	1.994874	3.60265	0.788612	3.836927	2.469485	2.233182	2.644462	1.779064	3.643883	3.550908	3.93575	2.06071
14	11	PMP	*	0	serous borc	59	borderline	boardline	42.8	CA	1.688057	3.991001	1.181827	3.729107	2.565115	2.673183	3.072414	2.100621	3.941649	3.556618	4.124222	2.414724
15	110	PMP	*	*	serous cyst	52	benign	benign	‡	*	1.563689	4.168147	0.946675	3.935105	2.651534	2.431707	3.165261	2.031309	3.995825	3.829709	4.083563	1.550175
16	111	PMP	*	*	serous cyst	49	benign	benign	‡	*	1.363483	3.90298	1.782944	4.157836	2.754145	3.169608	2.620071	2.301248	3.675944	3.425196	3.625964	1.457101
17	112	PMP	III	3	mixed with	57	III/IV	invasive	20.9	BBC	2.222003	4.132299	1.297726	3.9602	2.756132	2.489379	3.073251	2.129538	4.068948	3.7833	4.117629	1.927487
18	113	PMP	I	1	endometric	65	I/II	invasive	‡	*	1.809257	3.952609	1.711772	4.089466	2.725685	2.877331	2.884553	2.119838	3.813429	3.572154	3.780232	1.286774
19	114	PMP	II	3	serous ader	45	I/II	invasive	19.4	BBC	1.927998	3.980852	1.208797	4.290834	2.750331	2.674247	3.087048	1.999373	3.948632	3.653166	4.049903	1.693467
20	115	PMP	I	2	serous ader	52	I/II	invasive	‡	*	1.909338	3.996287	1.354376	3.912793	2.603325	2.676512	3.122608	2.094614	3.910578	3.611819	3.953515	1.550438
21	116	PMP	I	1	serous ader	73	I/II	invasive	128	CA	1.957366	4.068748	1.1566	4.383122	2.857974	2.56622	3.009838	2.220269	4.002625	3.788261	4.114454	1.103446
22	117	PMP	*	0	serous borc	62	borderline	boardline	42	CA	2.079113	3.535934	1.352736	4.341791	2.588751	2.624019	2.580081	1.751645	3.447535	3.508015	3.836654	1.590392
23	118	PMP	*	0	serous borc	61	borderline	boardline	‡	*	1.895986	3.606237	1.181827	4.3596	2.546303	2.559551	2.688493	1.422525	3.418946	3.522385	3.935804	0.890676
24	119	PMP	II	3	endometric	46	I/II	invasive	371	CA	1.688492	3.566177	1.232757	3.936407	2.397266	2.554341	2.315029	1.847004	3.441327	3.2092	3.419701	1.226399
25	12	PMP	II	2	clear cell a	59	I/II	invasive	125	CA	1.605447	4.19532	0.990552	3.545807	2.603325	2.637089	3.146248	1.842431	4.249165	3.867669	4.278484	1.847784
26	120	PMP	*	*	endometric	55	benign	benign	4.2	BBC	1.875833	3.991119	1.475771	3.95212	2.696676	2.722695	3.01667	2.270918	3.941912	3.565202	3.794142	1.825189
27	13	PMP	*	*	serous cyst	70	benign	benign	12	BBC	1.775221	4.080932	1.402959	3.767577	2.569941	2.738647	3.124515	1.903352	4.005133	3.709852	4.01349	2.22906
28	14	PMP	III	3	endometric	57	III/IV	invasive	382.2	CA	2.119942	3.069252	1.183467	4.340032	2.234647	3.130066	2.383625	1.35761	3.090367	3.026664	3.706546	-0.15015
29	15	PMP	*	0	serous borc	56	borderline	boardline	450.3	CA	1.698685	4.193465	1.12636	3.778023	2.717338	2.643231	3.125647	1.824495	4.197283	3.847564	4.243657	1.736994
30	16	PMP	III	3	endometric	63	III/IV	invasive	1002	CA	1.677587	4.101587	1.1566	3.949511	2.732478	2.834927	3.087048	1.914381	4.017649	3.520597	4.141287	2.010683
31	17	PMP	*	0	serous borc	69	borderline	boardline	388	CA	1.57792	3.547338	1.386842	4.545	2.492773	3.176411	2.909948	1.773715	3.1007	2.960178	3.51765	0.577276
32	18	PMP	*	0	serous borc	50	borderline	boardline	118	CA	1.719167	3.851681	1.433534	3.293219	2.31643	2.790763	2.815361	1.555404	3.829598	3.644216	3.901004	2.358547
33	19	PMP	III	2	endometric	55	III/IV	invasive	‡	*	1.809257	4.256018	1.317425	3.766146	2.822789	2.728867	3.114951	1.83781	4.209215	3.872307	4.107927	2.144539
34	2	PMP	*	0	serous borc	58	borderline	boardline	763	CA	1.939559	3.975204	1.181827	3.725716	2.693751	2.67546	2.970325	1.344445	4.000097	3.790741	4.235418	2.363795
35	20	PMP	II	3	serous ader	74	I/II	invasive	9	BBC	2.044176	4.305862	1.463774	4.327389	3.032641	2.897492	3.27145	2.161899	4.144658	3.71961	4.333168	1.863838

The key point what we got

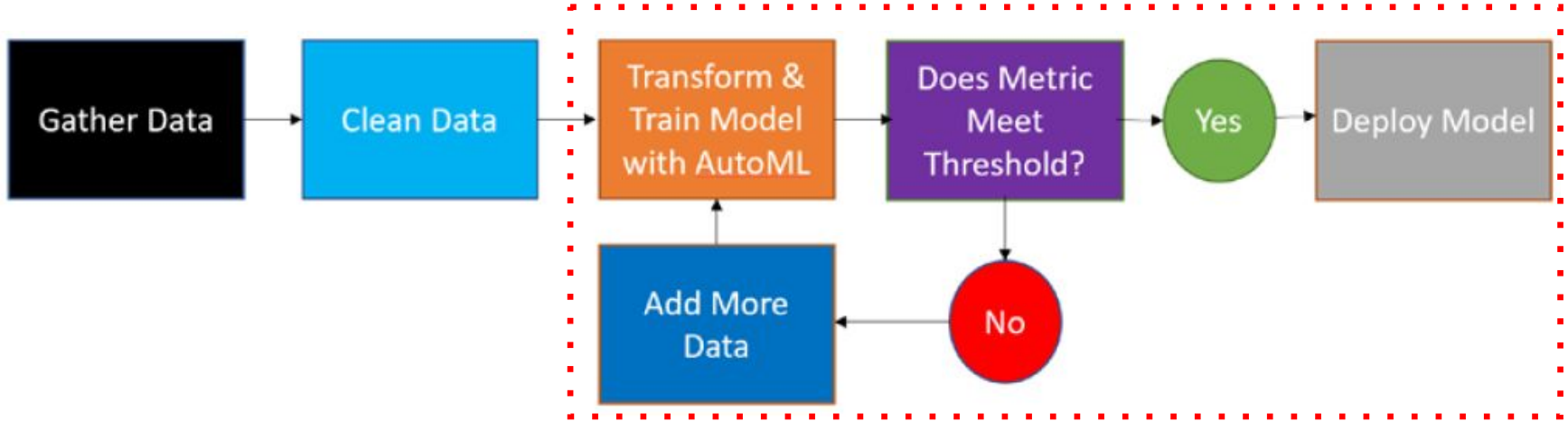
- Why the ML algorithms without boosting ensemble models, such as XGBoost ??
  - In recent years, XGBoost is a popular tool among a lot of campaigns due to its high performance.
- How to deal with hyperparameter tuning ??
  - Some ML algorithm must tune hyperparameters for performing well, e.g. XGBoost .
- If there is a automatic way to select feasure??

Traditional ML vs AutoML

Traditional Machine Learning



AutoML



rminer package in r

- new automated machine learning (AutoML)
  - Versions: 1.4.6 / 1.4.5 / 1.4.4
  - 14 classification and 15 regression methods
- Multi-class Classification
  - conditional inference tree
  - generalized linear model (GLM) with elastic net regularization
  - decision tree
  - k-nearest neighbor
  - support vector machine
  - least squares support vector machine
  - multilayer perceptron with one hidden layer
  - random forest
  - XGBoost
  - Bagging
  - adaboost.M1 method
  - linear discriminant analysis
  - multinomial logistic regression

The tricky thing about data

Table 2. Demographics of patients after stratified random sampling into training and testing sets.

	Training (n = 135)	Testing (n = 44)	p-value
Age, years, median (SD) *	**56 (8.1)**	**56 (8.3)**	**1.0**
CA-125, units/ml, median (SD) *	126.5 (1193.5)	105.6 (577.8)	0.91
Pathology, n (%)†			1.0
Control	11 (8.1)	4 (9.1)	
Benign lesions	34 (25.2)	11 (25.0)	
Borderline tumors	16 (11.9)	5 (11.4)	
Stage I/II invasive cancers	41 (30.4)	12 (27.3)	
Stage III/IV invasive cancers	33 (24.4)	12 (27.3)	

Stage III/IV: 45

Stage III/IV: 46 &
3 patients coding borderline

A	B	C	D	E
ID	Study	Stage	Grade	Histology
e1012	ERA	III	0	serous borderline
e1027	ERA	III	0	serous borderline
e1048	ERA	III	0	serous borderline

Grade =0 & Histology written as “borderline” means that those case were checked by examination in the pathology lab

All tool we used

- R
  - openxlsx
  - rminer
  - sampling
  - caret
- Python
  - klearn
  - sklego
  - pgmpy.models
  - pandas
  - numpy

The same feasures in original models

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

What we found[1] ?

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

Our result

	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

Table 4. Performance of the eleven statistical models on the testing set by variable selection method. Results are shown for the testing set.

	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)

What we found[2] ?

- boosting ensemble models may be another good choice .

Model	classification accuracy rate	weighted True Positive rate	weighted F1 score
conditional inference tree(ctree)	50	50	38.3
generalized linear model (GLM) with lasso or elasticnet regularization(cv.glmnet)	54.5	54.5	40.6
decision tree(dt)	54.5	54.5	52.1
k-nearest neighbor(knn)	50	50	50.5
support vector machine(svm)	52.3	52.3	35.9
least squares support vector machine(lssvm)	52.3	56.8	59.1
Neural Network(multilayer perceptron, mlp)	47.7	43.2	50.5
random Forest	52.3	56.8	45.7
xgboost	56.8	56.8	53.1
bagging	56.8	56.8	49.7
adaboost.M1 method(boosting)	56.8	61.4	53.1
linear discriminant analysis(lda)	43.2	43.2	42.6
logistic regression(multinomial)	47.7	47.7	47.2
naiveBayes	43.2	43.2	44.2

What we found[3] ?

- We tried to use Recursive Feature Elimination to select miRNA automatically
  - The result are inconsistent

Reproducibility

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

- 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


Thank You !