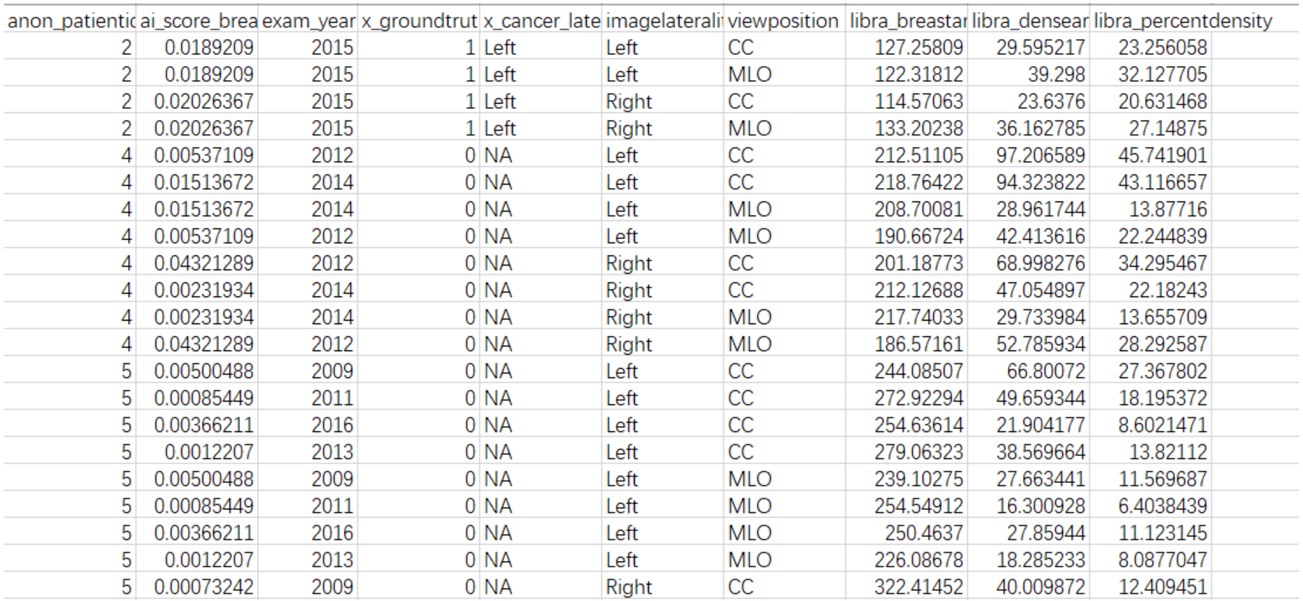
Analysis of AI scores from the mammography exam

Reporter: Rudan Xiao

Date: 11/04/2022

- Data Description and Statistical Analysis
  - Data description
  - Data distribution & correlation analysis
  - Nonparametric tests
- Evaluation Index & AI score Related Features
- Methods of Classification
- Screening for Potential Cancer
- Conclusion

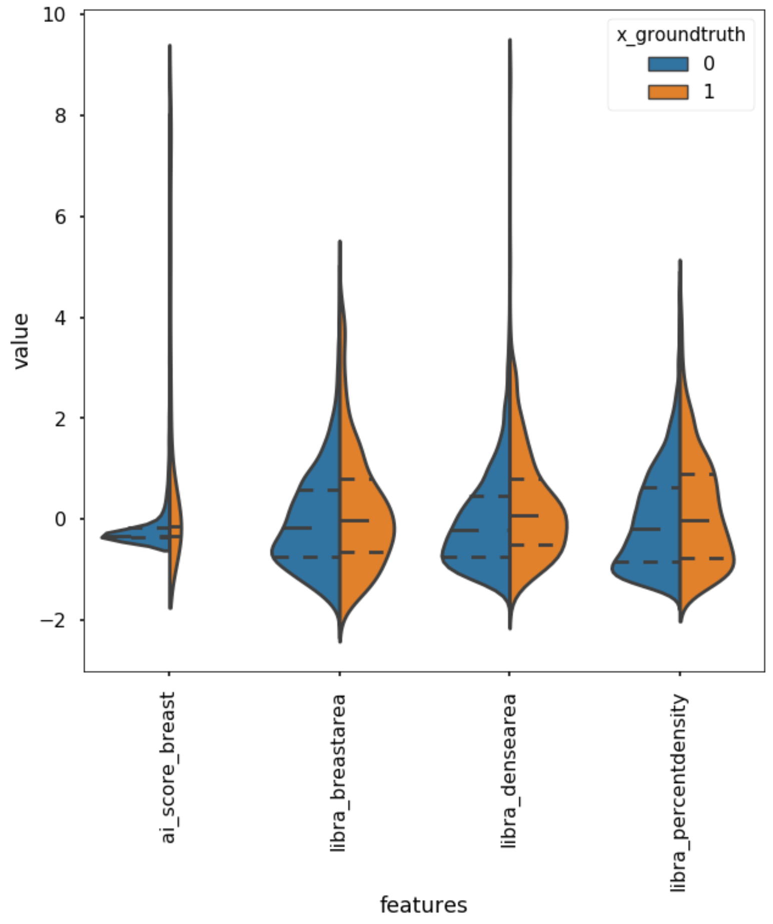


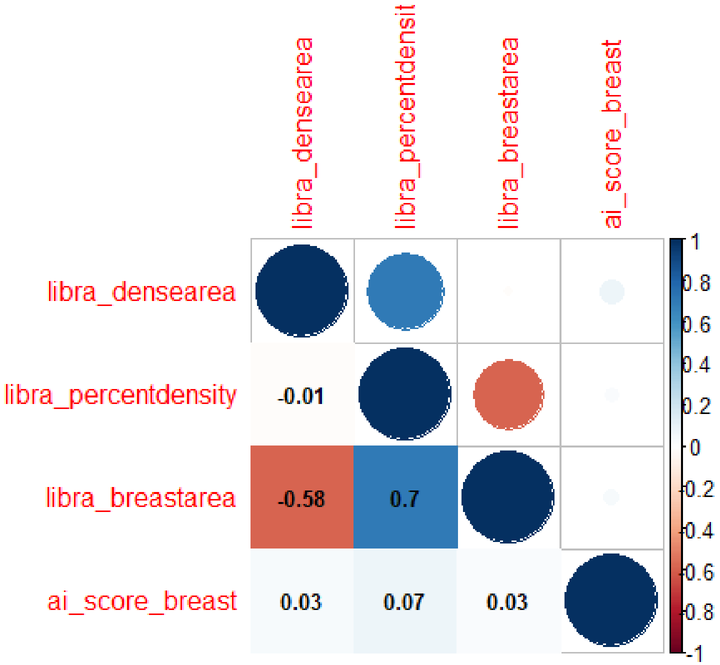
anon_patientic	ai_score_brea	exam_year	x_groundtrut	x_cancer_late	imagelaterali	viewposition	libra_breastar	libra_densear	libra_percentdensity
2	0.0189209	2015	1	Left	Left	CC	127.25809	29.595217	23.256058
2	0.0189209	2015	1	Left	Left	MLO	122.31812	39.298	32.127705
2	0.02026367	2015	1	Left	Right	CC	114.57063	23.6376	20.631468
2	0.02026367	2015	1	Left	Right	MLO	133.20238	36.162785	27.14875
4	0.00537109	2012	0	NA	Left	CC	212.51105	97.206589	45.741901
4	0.01513672	2014	0	NA	Left	CC	218.76422	94.323822	43.116657
4	0.01513672	2014	0	NA	Left	MLO	208.70081	28.961744	13.87716
4	0.00537109	2012	0	NA	Left	MLO	190.66724	42.413616	22.244839
4	0.04321289	2012	0	NA	Right	CC	201.18773	68.998276	34.295467
4	0.00231934	2014	0	NA	Right	CC	212.12688	47.054897	22.18243
4	0.00231934	2014	0	NA	Right	MLO	217.74033	29.733984	13.655709
4	0.04321289	2012	0	NA	Right	MLO	186.57161	52.785934	28.292587
5	0.00500488	2009	0	NA	Left	CC	244.08507	66.80072	27.367802
5	0.00085449	2011	0	NA	Left	CC	272.92294	49.659344	18.195372
5	0.00366211	2016	0	NA	Left	CC	254.63614	21.904177	8.6021471
5	0.0012207	2013	0	NA	Left	CC	279.06323	38.569664	13.82112
5	0.00500488	2009	0	NA	Left	MLO	239.10275	27.663441	11.569687
5	0.00085449	2011	0	NA	Left	MLO	254.54912	16.300928	6.4038439
5	0.00366211	2016	0	NA	Left	MLO	250.4637	27.85944	11.123145
5	0.0012207	2013	0	NA	Left	MLO	226.08678	18.285233	8.0877047
5	0.00073242	2009	0	NA	Right	CC	322.41452	40.009872	12.409451

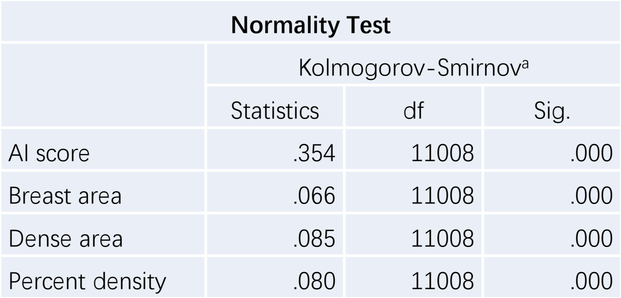
Number of samples

	cancer	healthy
patients	57	721
Ai score	632	10376
Cancer in left laterality	332 (33)	___
Cancer in right laterality	300 (24)	___

- ➢ Which AI score is most relevant to the cancer diagnosis?
- ➢ Which method works best for classifying cancer and healthy?
- ➢ Which healthy person should be recalled for the next diagnosis?





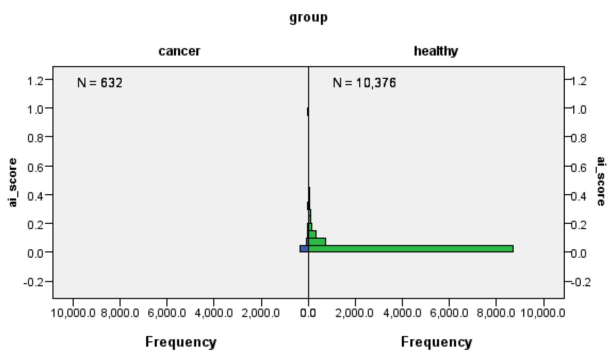


Normality Test			
	Kolmogorov-Smirnov[a]		
	Statistics	df	Sig.
AI score	.354	11008	.000
Breast area	.066	11008	.000
Dense area	.085	11008	.000
Percent density	.080	11008	.000

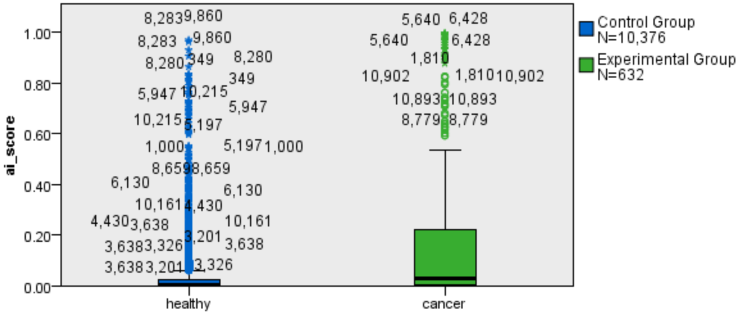
Data do **not** follow a **normal** distribution and **unbalanced**

AI score is more related with **breast area**.

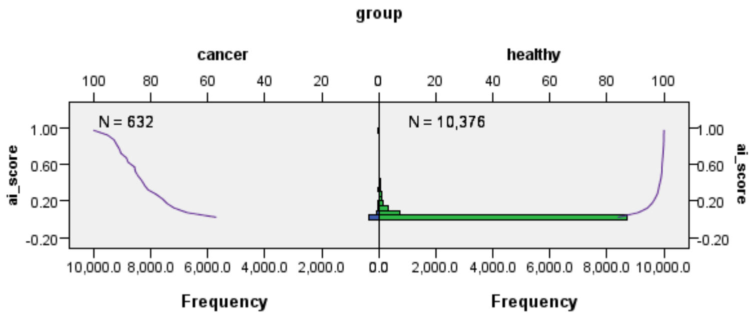
Independent-Samples Wald-Wolfowitz Runs Test

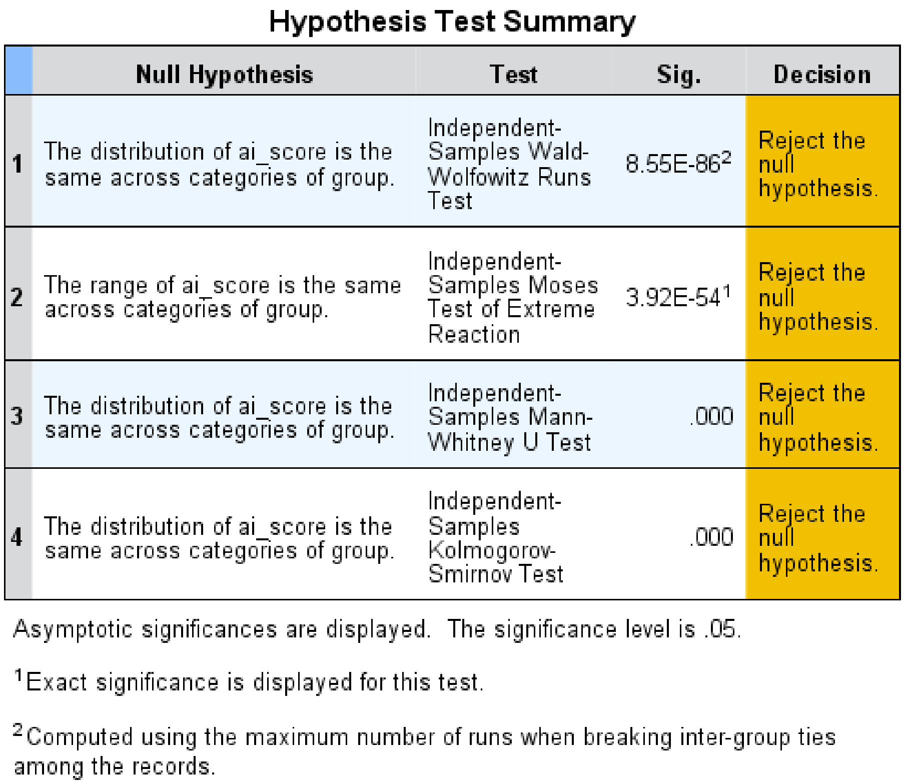


Independent-Samples Moses Test of Extreme Reaction



Independent-Samples Kolmogorov-Smirnov Test





Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ai_score is the same across categories of group.	Independent-Samples Wald-Wolfowitz Runs Test	8.55E-86[2]	Reject the null hypothesis.
2	The range of ai_score is the same across categories of group.	Independent-Samples Moses Test of Extreme Reaction	3.92E-54[1]	Reject the null hypothesis.
3	The distribution of ai_score is the same across categories of group.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
4	The distribution of ai_score is the same across categories of group.	Independent-Samples Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

[1] Exact significance is displayed for this test.

[2] Computed using the maximum number of runs when breaking inter-group ties among the records.

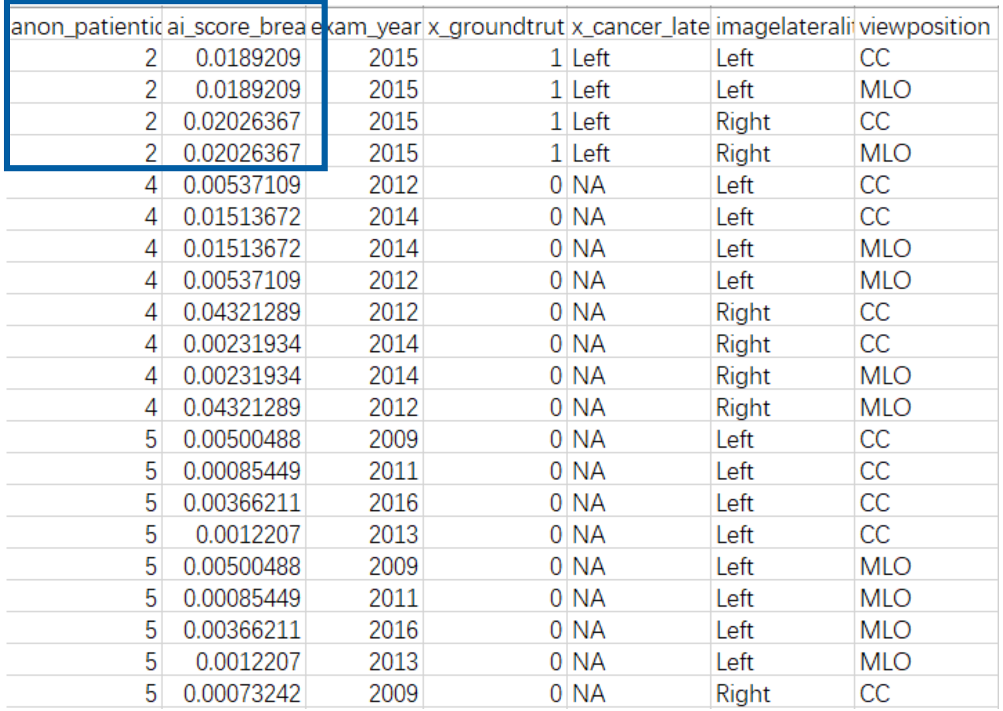
The **difference in AI-score** between **cancer** patients (ground truth is 1) and the **healthy** patients (ground truth is 0) are **significant** under all test methods.

- Data Description and Statistical Analysis
- Evaluation Index & AI score Related Features
    - Evaluation index
    - AI score related features
- Methods of Classification
- Screening for Potential Cancer
- Conclusion

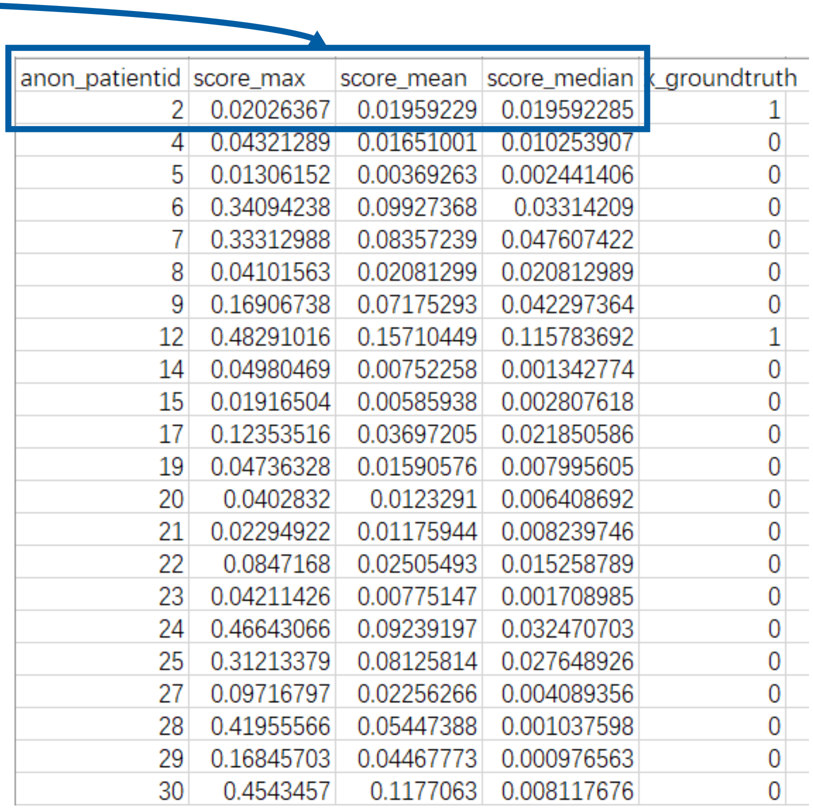
Evaluation index of classification

index	Notes	Formula
sensitivity (recall, true positive rate TPR)	the ability of a test to correctly identify patients with a disease	TP/(TP+FN)
specificity (selectivity, true negative rate (TNR))	the ability of a test to correctly identify people without the disease	TN/(TN+FP)
accuracy	Ratio of correct predictions to total predictions	(TP+TN)/(TP+TN+FP+FN)
precision (positive predictive value (PPV))	Ratio of true positives to total predicted positives	TP/(TP+FP)
f1_score	the harmonic mean of precision and sensitivity	2TP/(2TP+FP+FN)

Calculate the **max**, **mean**, and **median of AI score** by a group of patient ID

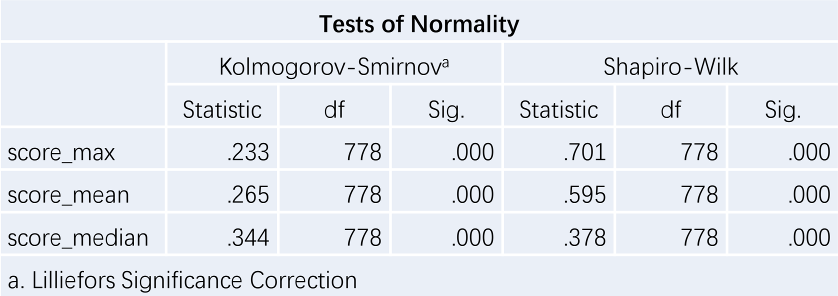


anon_patientic	ai_score_brea	exam_year	x_groundtrut	x_cancer_late	imagelaterali	viewposition
2	0.0189209	2015	1	Left	Left	CC
2	0.0189209	2015	1	Left	Left	MLO
2	0.02026367	2015	1	Left	Right	CC
2	0.02026367	2015	1	Left	Right	MLO
4	0.00537109	2012	0	NA	Left	CC
4	0.01513672	2014	0	NA	Left	CC
4	0.01513672	2014	0	NA	Left	MLO
4	0.00537109	2012	0	NA	Left	MLO
4	0.04321289	2012	0	NA	Right	CC
4	0.00231934	2014	0	NA	Right	CC
4	0.00231934	2014	0	NA	Right	MLO
4	0.04321289	2012	0	NA	Right	MLO
5	0.00500488	2009	0	NA	Left	CC
5	0.00085449	2011	0	NA	Left	CC
5	0.00366211	2016	0	NA	Left	CC
5	0.0012207	2013	0	NA	Left	CC
5	0.00500488	2009	0	NA	Left	MLO
5	0.00085449	2011	0	NA	Left	MLO
5	0.00366211	2016	0	NA	Left	MLO
5	0.0012207	2013	0	NA	Left	MLO
5	0.00073242	2009	0	NA	Right	CC



anon_patientid	score_max	score_mean	score_median	x_groundtruth
2	0.02026367	0.01959229	0.019592285	1
4	0.04321289	0.01651001	0.010253907	0
5	0.01306152	0.00369263	0.002441406	0
6	0.34094238	0.09927368	0.03314209	0
7	0.33312988	0.08357239	0.047607422	0
8	0.04101563	0.02081299	0.020812989	0
9	0.16906738	0.07175293	0.042297364	0
12	0.48291016	0.15710449	0.115783692	1
14	0.04980469	0.00752258	0.001342774	0
15	0.01916504	0.00585938	0.002807618	0
17	0.12353516	0.03697205	0.021850586	0
19	0.04736328	0.01590576	0.007995605	0
20	0.0402832	0.0123291	0.006408692	0
21	0.02294922	0.01175944	0.008239746	0
22	0.0847168	0.02505493	0.015258789	0
23	0.04211426	0.00775147	0.001708985	0
24	0.46643066	0.09239197	0.032470703	0
25	0.31213379	0.08125814	0.027648926	0
27	0.09716797	0.02256266	0.004089356	0
28	0.41955566	0.05447388	0.001037598	0
29	0.16845703	0.04467773	0.000976563	0
30	0.4543457	0.1177063	0.008117676	0

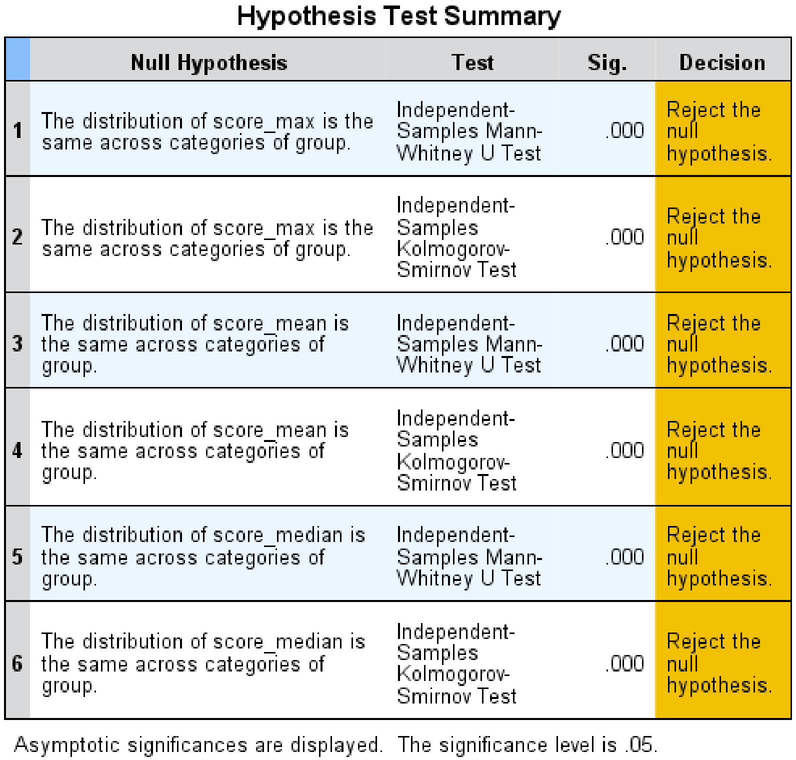
Nonparametric Tests



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
score_max	.233	778	.000	.701	778	.000
score_mean	.265	778	.000	.595	778	.000
score_median	.344	778	.000	.378	778	.000

a. Lilliefors Significance Correction

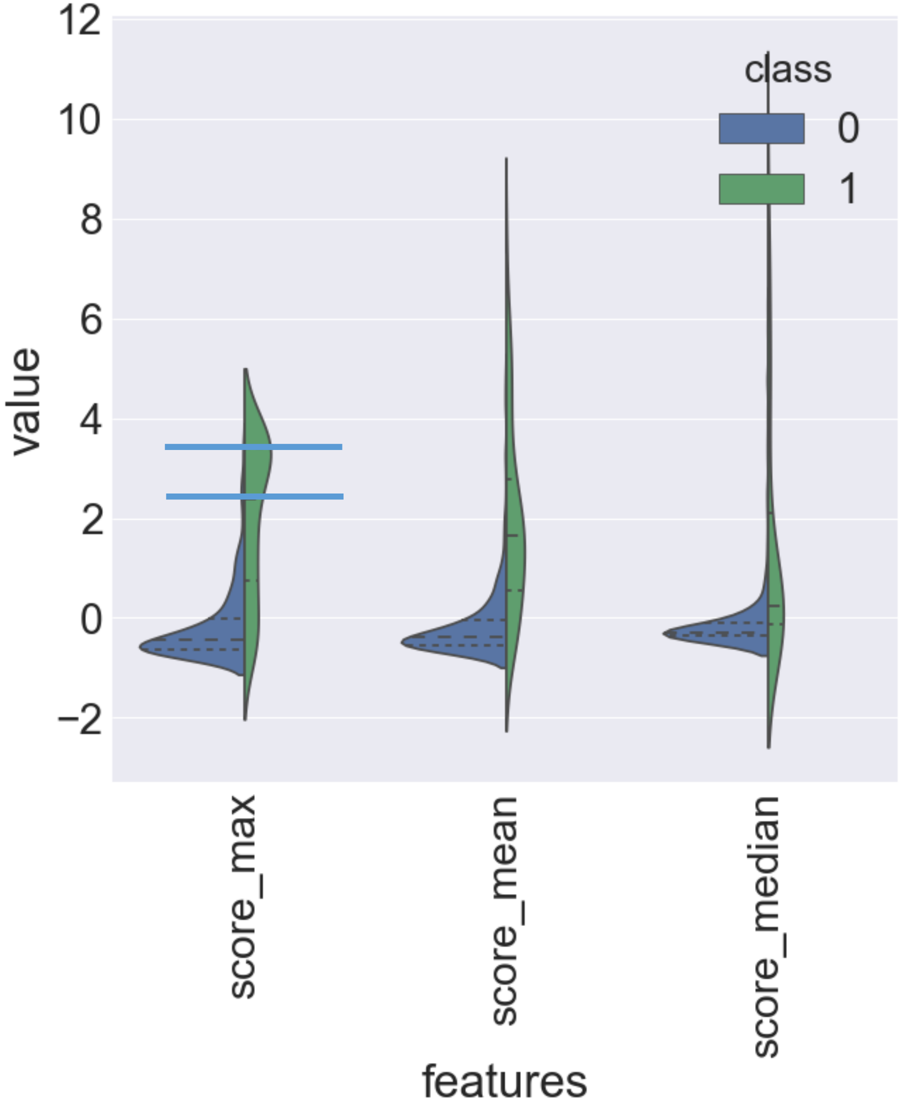


Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of score_max is the same across categories of group.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
2	The distribution of score_max is the same across categories of group.	Independent-Samples Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.
3	The distribution of score_mean is the same across categories of group.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
4	The distribution of score_mean is the same across categories of group.	Independent-Samples Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.
5	The distribution of score_median is the same across categories of group.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
6	The distribution of score_median is the same across categories of group.	Independent-Samples Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.

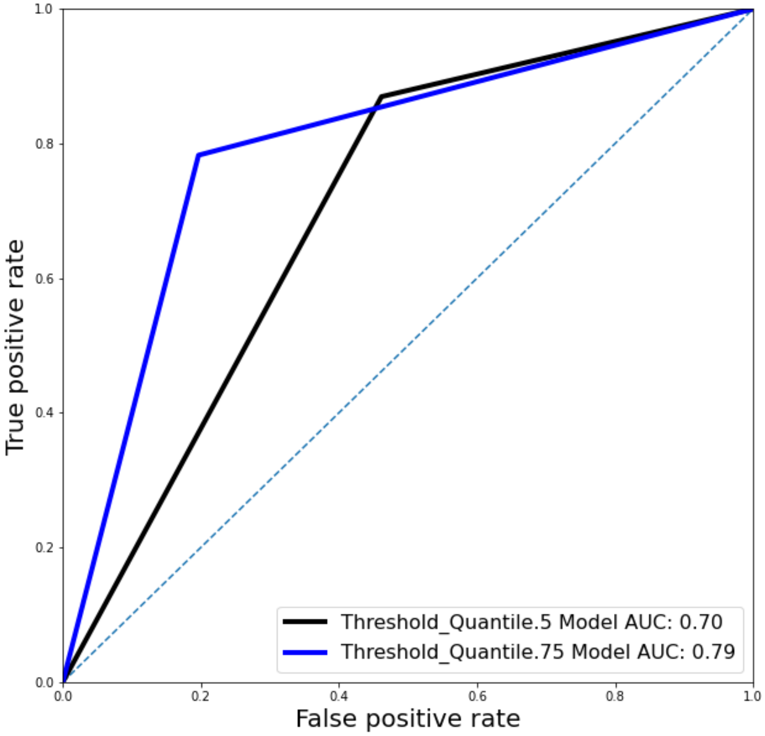
Asymptotic significances are displayed. The significance level is .05.

The **difference** in **three features (score_max, score_mean, and score_median)** between **cancer** patients (ground truth is 1) and the **healthy** patients (ground truth is 0) are **significant** under all test methods



From the **features distribution** of class 0 (healthy) and class 1 (cancer), **score_max** seems to be easier to distinguish **by quartile threshold.**

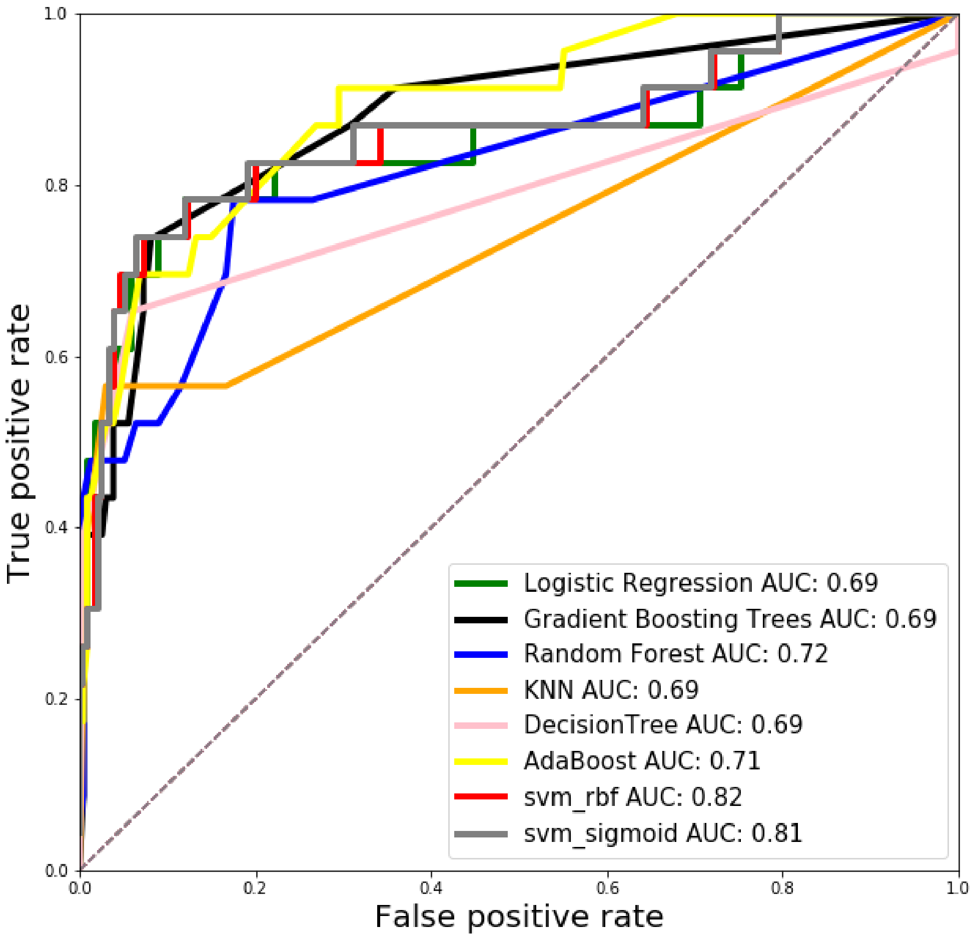
- Data Description and Statistical Analysis
- Evaluation Index & AI score Related Features
- Methods of Classification
  - Threshold
  - Traditional machine learning
  - Multi-Layer perceptron (MLP)
- Screening for Potential Cancer
- Conclusion



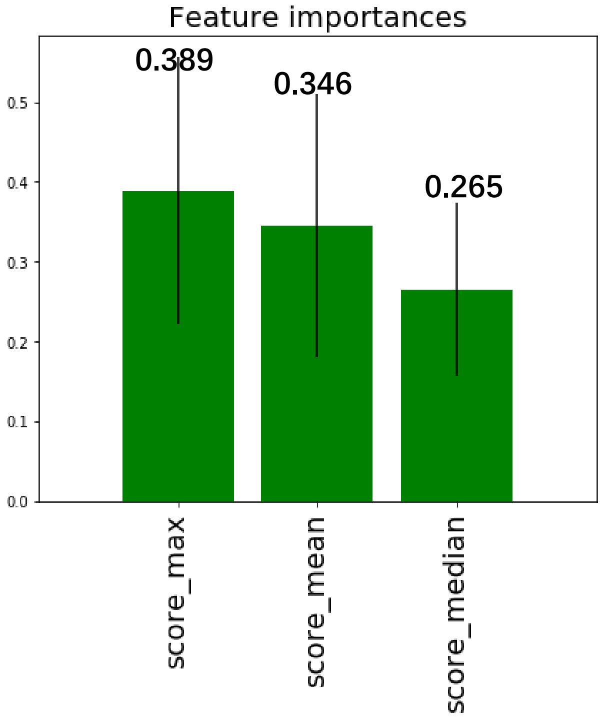
Receiver Operating Characteristic curve (ROC)

Evaluation Result

index	Threshold (Quartile=0.5)	Threshold (Quartile=0.75)
sensitivity	**0.8772**	0.7895
specificity	0.5298	**0.7920**
accuracy	0.5553	**0.7918**
precision	0.1285	**0.2308**
f1_score	0.2242	**0.3571**
AUC	0.70	**0.79**



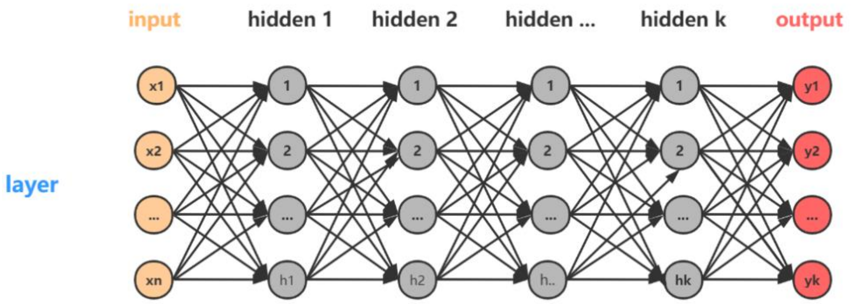
Receiver Operating Characteristic curve (ROC)



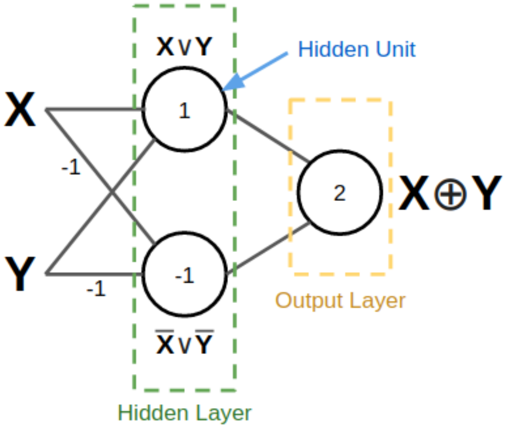
- **SVM_rbf** performs **best** among all the traditional machine learning methods of classification.
- **Score_max** performs **best** among all the features.

The evaluation result of all the traditional machine learning models

	sensitivity	specificity	accuracy	precision	f1_score	AUC
Logistic Regression	0.3478	0.9915	0.9339	0.8	0.4848	0.69
Gradient Boosting Trees	0.3913	**0.9957**	0.9416	0.9	0.5455	0.69
Random Forest	0.4348	**0.9957**	**0.9455**	**0.9091**	0.5882	0.72
KNN	0.4348	0.9786	0.93	0.6667	0.5263	0.69
DecisionTree	0.3913	**0.9957**	0.9416	0.9	0.5455	0.69
AdaBoost	0.3913	0.9915	0.9377	0.8182	0.5294	0.71
svm_rbf	**0.6957**	0.9359	0.9144	0.5161	**0.5926**	**0.82**
svm_sigmoid	0.6087	0.6087	0.93	0.6087	0.6087	0.81



MLP is a kind of **neural network**



With a **layered** structure, we can obtain the **intermediate results** and use them **as the inputs to the next layer**. We call this structure the multi-layer perceptron (MLP)

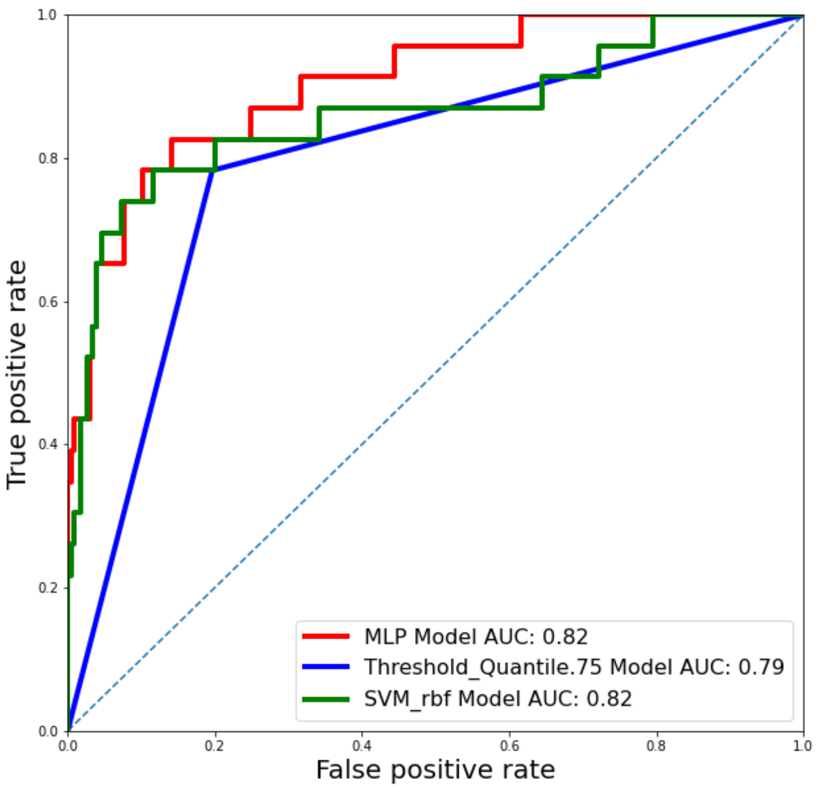
$$L(w) = \sum_{x_i \in D} -y_i(w^T x_i + b),$$

$$D = \{x_i \mid y_i(w^T x_i + b) < 0\}$$

Loss Function

Evaluation Result

index	Threshold (Quartile=0.5)
sensitivity	0.4348
specificity	0.9872
accuracy	0.9377
precision	0.7692
f1_score	0.5556
AUC	0.82

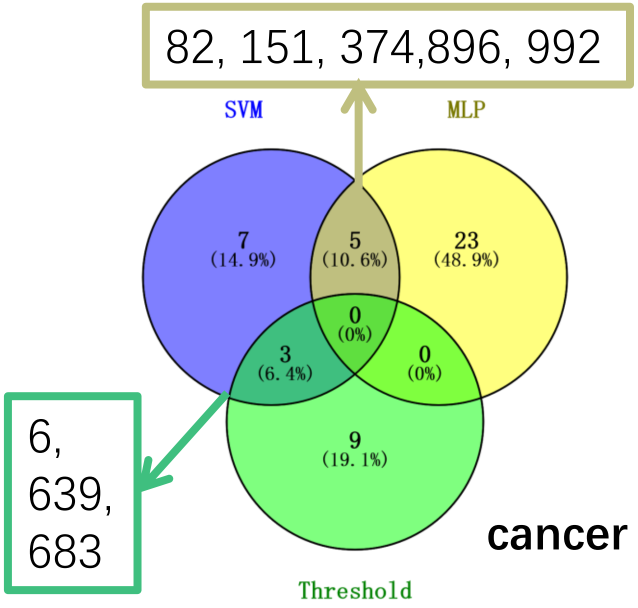


Receiver Operating Characteristic curve (ROC)

Significant test in the difference of AUC by DeLong

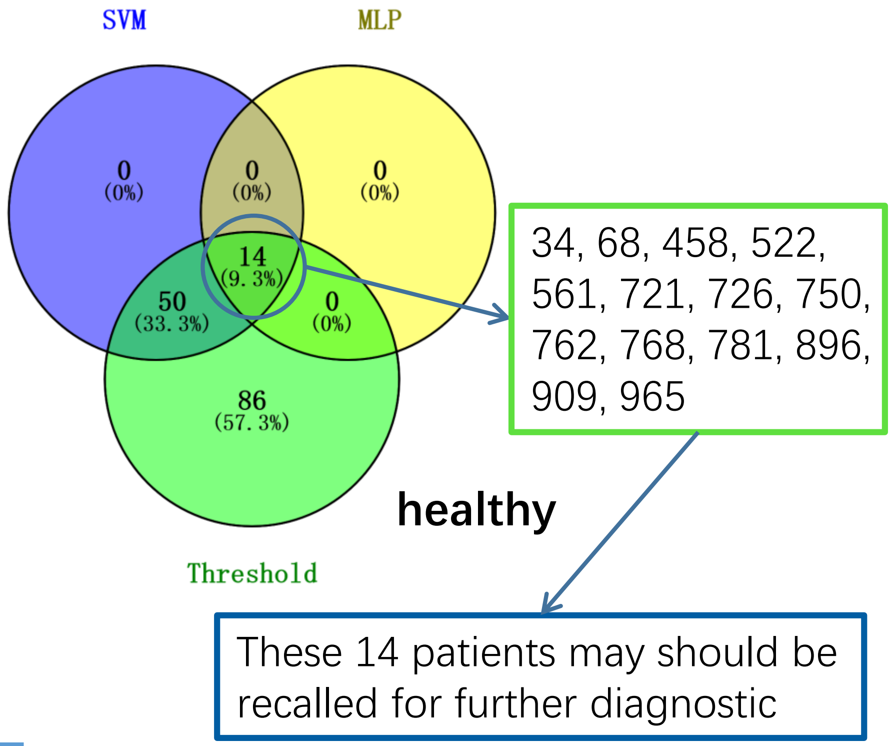
	P-Value (0.05)
MLP V.S. **SVM_rbf**	0.2377
Threshold V.S. **SVM_rbf**	0.1098
MLP V.S. **Threshold**	**0.0067**

- Data Description and Statistical Analysis
- Evaluation Index & AI score Related Features
- Methods of Classification
- Screening for Potential Cancer
- Conclusion



82, 151, 374,896, 992

6, 639, 683



34, 68, 458, 522, 561, 721, 726, 750, 762, 768, 781, 896, 909, 965

These 14 patients may should be recalled for further diagnostic

https://bioinfogp.cnb.csic.es/tools/venny/index.html

Number of wrong classify patients

	MLP	SVM_rbf	Threshold (Quartile=0.75)
Cancer (x_groundtruth=1)	28	15	12
Healthy (x_groundtruth=0)	14	64	150

- The three classifiers are **more consistent** in the **classification of healthy** people (possibly due to data imbalanced).
- **14** patients **misclassified** as cancer in **all three classifiers** may need to be **recalled**.

- Data Description and Statistical Analysis
- Evaluation Index & AI score Related Features
- Methods of Classification
- Screening for Potential Cancer
- Conclusion

1. **MLP** performs **best** among all **classifiers**.

2. **Traditional machine learning** models are more **explanatory** and can explain the **contribution** of each **feature**.

3. The **distribution of features** data and the **features rank** results of machine learning both show that the **score_max** has the **largest contribution**.

Thanks

- **AIR** (abnormal interpretation rate) seems to be commonly used for evaluation in **cohort studies**, combined with some clinical data.

- For the **time series analysis**, now the data seems does not include enough time information, for example, patients with groundtruth 1, what time is the **date of the first diagnosis**?

- **CC** and **MLO** have the **same AI score** value?

- Solve the problem of **unbalanced**.