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

Kevin M Elias, Wojciech Fendler, Konrad Stawiski, Stephen J Fiascone, Allison F Vitonis, Ross S Berkowitz, Gyorgy Frendl, Panagiotis Konstantinopoulos, Christopher P Crum, Magdalena Kedzierska, Daniel W Cramer, Dipanjan Chowdhury

陳先灝、盧佳妤、段寶鈞

課程：生物資訊概論與實務

2020/1/7

Outline

01 Introduction

02 Data

03 Preprocess

04 Models

05 Results

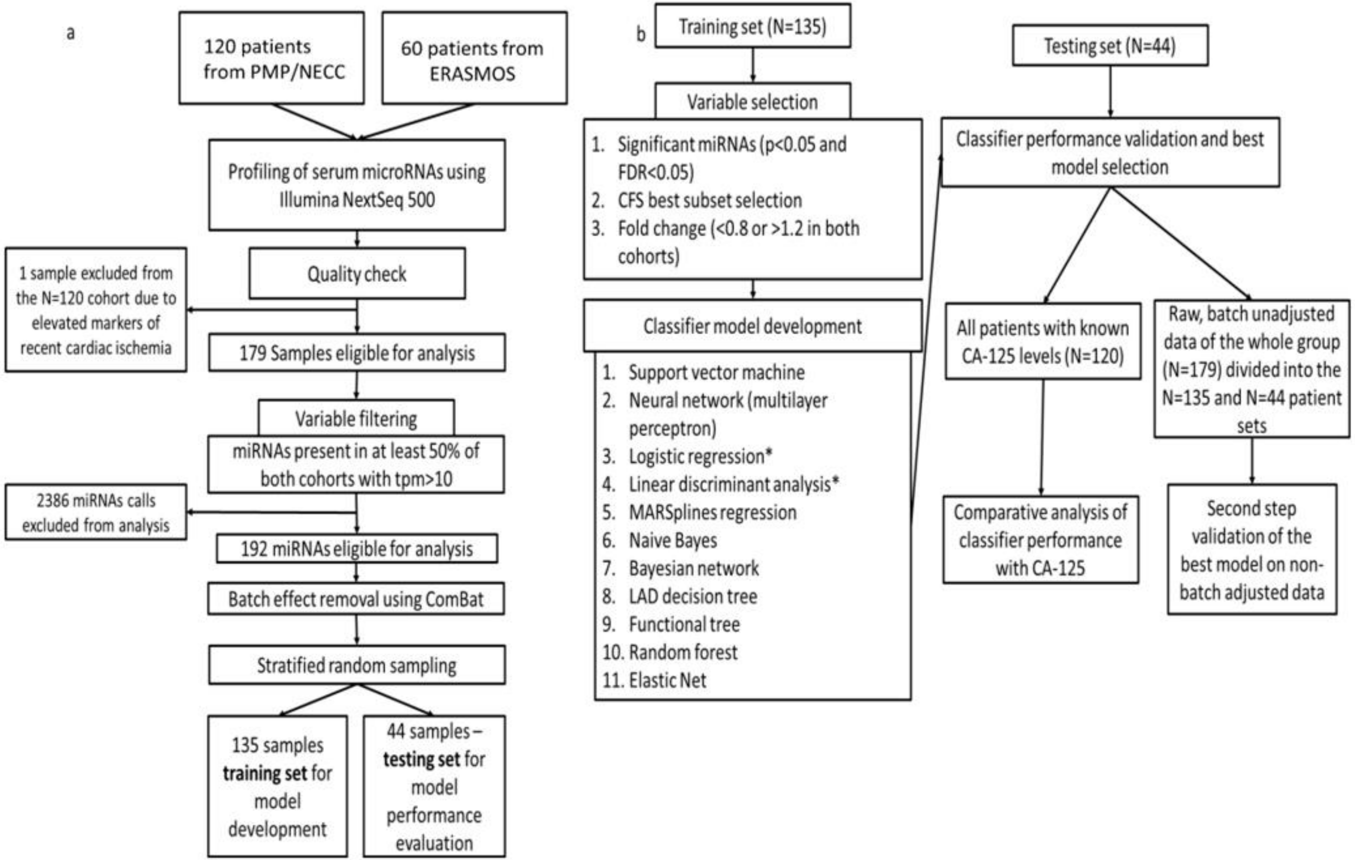
06 Others

Introduction

Introduction & Data Format

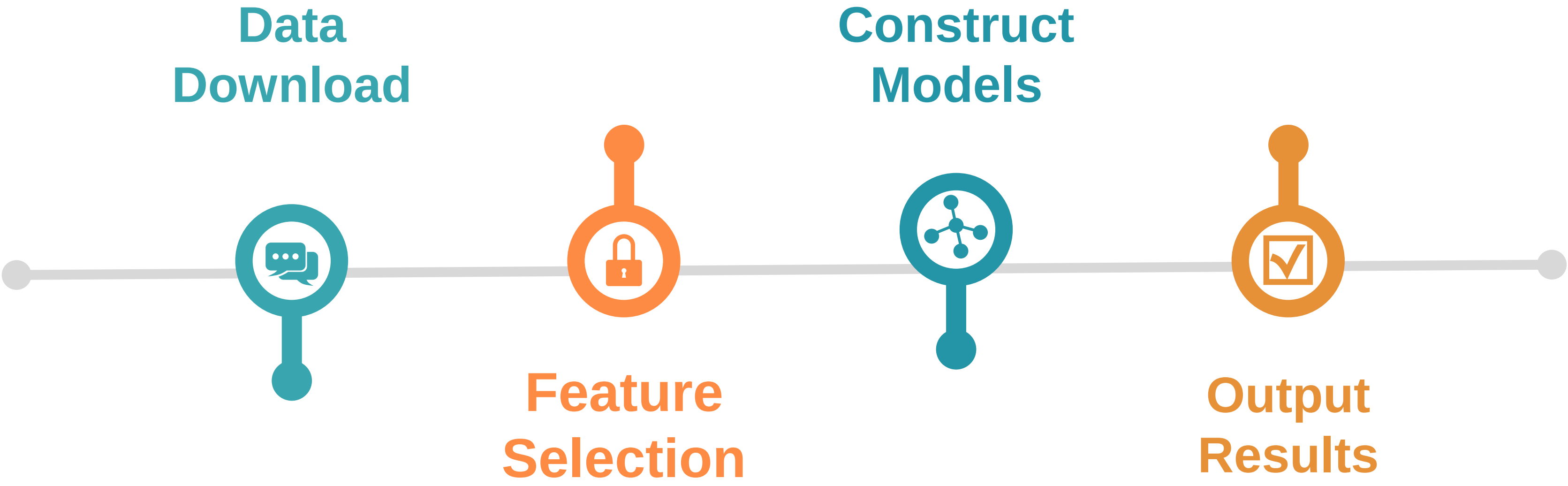
- **Motivation**
  - While recognizing epithelial ovarian cancer(EOC), protein (CA125) detection is not convenient enough, and have some shortcomings.
    - 1. microRNA can detect more rare transcripts in the blood through PCR.
    - 2. All microRNA are in the same unit measurement, which is easier to incorporate into multiplexed panels.
    - 3. miRNAs play a critical role in ovarian cancer biology, whereas the function of CA125 is unknown.
- **Data Property & Format**
  - microRNA with elements “A”, “T”, “C”, “G”.
  - Fastq format
    - We will use elements recorded in line 2, and maybe the quality values listed in line 4.
  - It can be read easily.

Flowchart of study design



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

Flowchart



Data Download

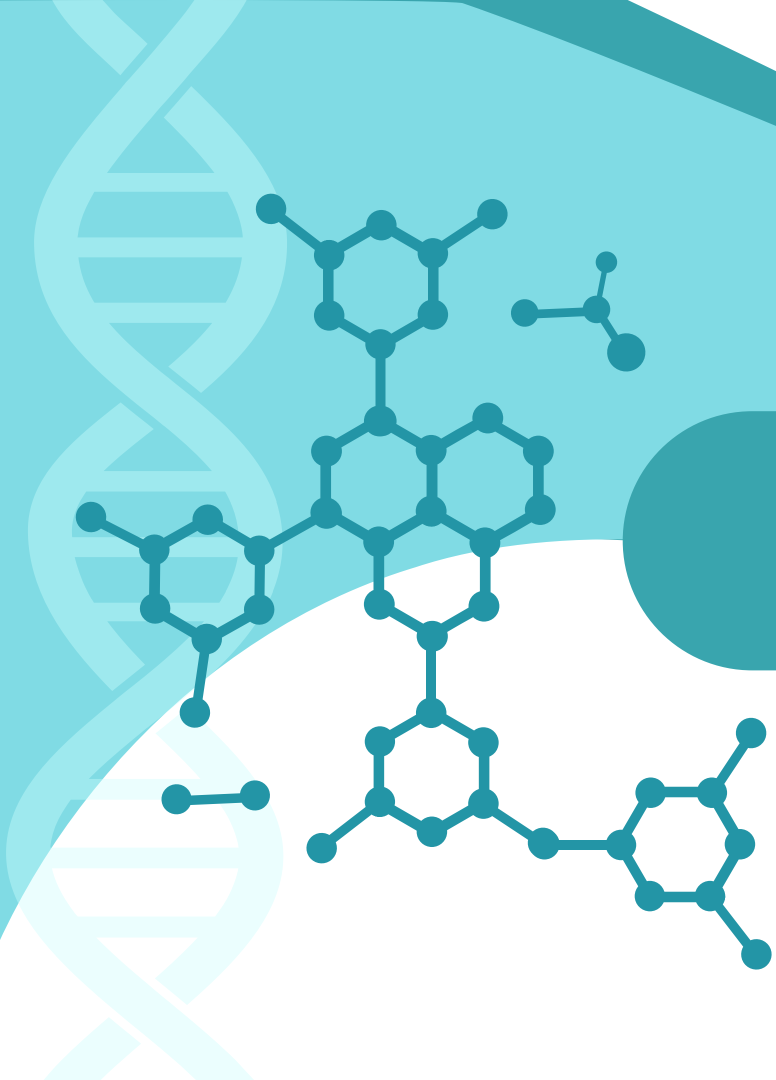
Feature Selection

Construct Models

Output Results

Tool

- **Paper**
  - STATISTICA Data Miner 12.5 (It’s not free.)
  - Weka 3.9.0
- **We want to use**
  - python
    - Numpy
    - Keras
    - pytorch
    - scikit-learn



Data

Data Download

- **Where to download?**
  - NCBI - GEO Accession viewer (GSE94533)
    Run Selector (PRJNA371423)
  - EBI - ENA Broswer (PRJNA371423)

- **How big in terms of GB? in terms of reads?**
  - sra - 48.78GB
  - fastq - 1.29GB * 204
  - xlsx - 4.89MB + 9.22MB

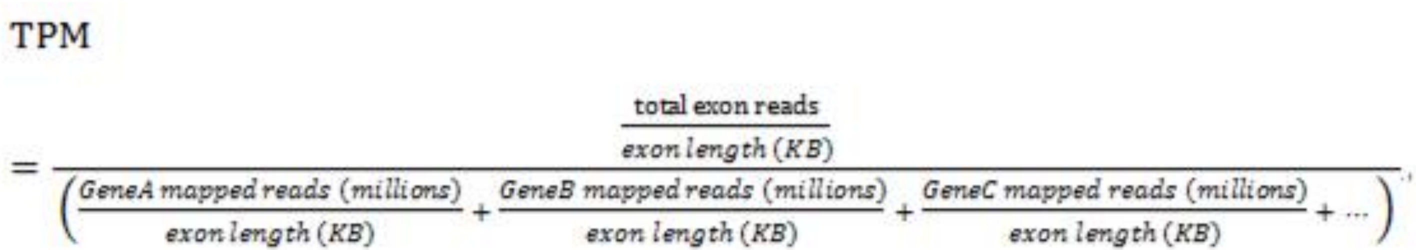
Preprocess

Preprocess

- **xlsx file - GSE94533_Processed_file_Cohort1.xlsx**
  GSE94533_Processed_file_Cohort2.xlsx
  - sample, 01_summary_all, mirna_rawcounts, mirna_tpm (Remap Name), smallrna_rawcounts, smallrna_tpm, putative_mirna
- **Run Selector - PRJNA371423**
  - Run, BioSample, AvgSpotLen, Bases, Bytes, diagnosis, Experiment, GEO_Accession, Histology, MBases, MBytes, Sample Name, stage
- **GEO Accession viewer (GSE94533)**
  - 180 of Samples (Sample Name/Remap Name)

Preprocess - Why mirna_tpm

- TPM (Transcript Per Million)
    - TPM is a unit used to measure expression in NGS experiments.
    - The number of reads for a particular miRNA is divided by the total number of mapped reads and multiplied by 1 million.



TPM

$$= \frac{\frac{\text{total exon reads}}{\text{exon length }(KB)}}{\left(\frac{\text{GeneA mapped reads (millions)}}{\text{exon length }(KB)} + \frac{\text{GeneB mapped reads (millions)}}{\text{exon length }(KB)} + \frac{\text{GeneC mapped reads (millions)}}{\text{exon length }(KB)} + \cdots\right)},$$

Feature Select

Significance Correlation Fold

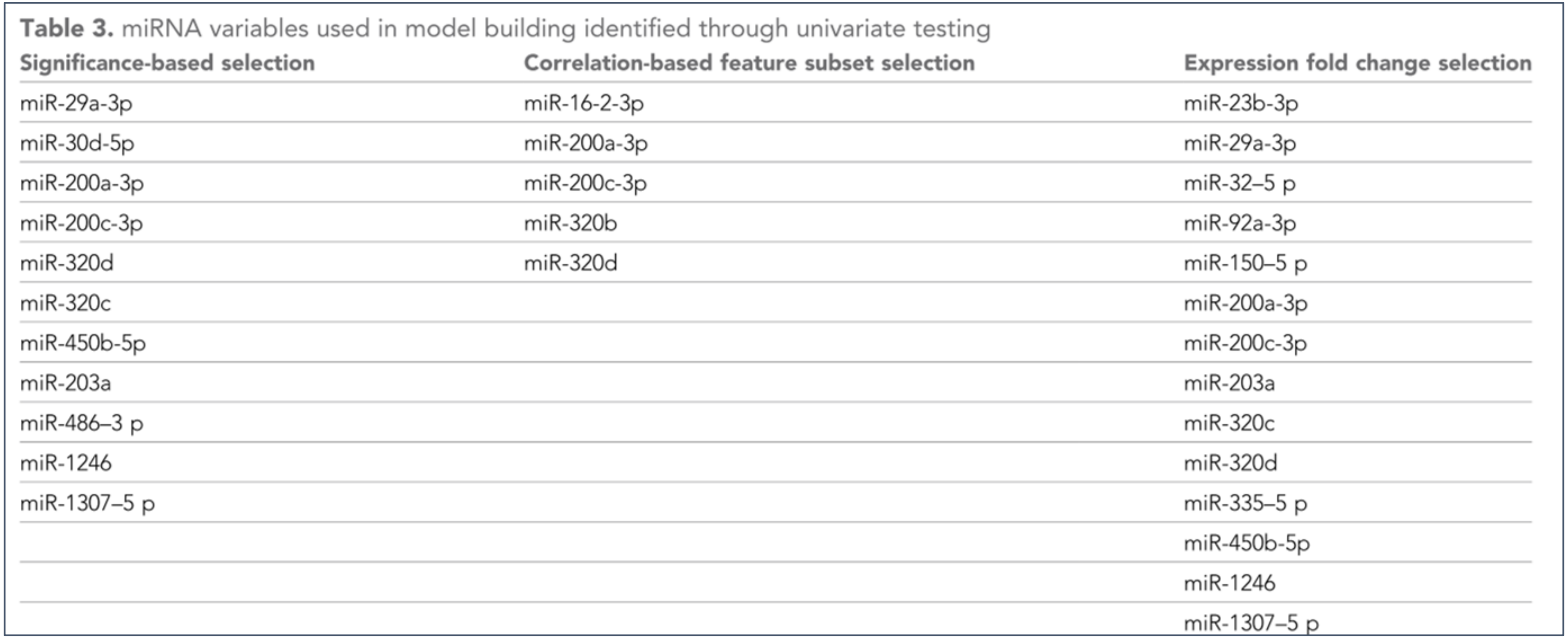
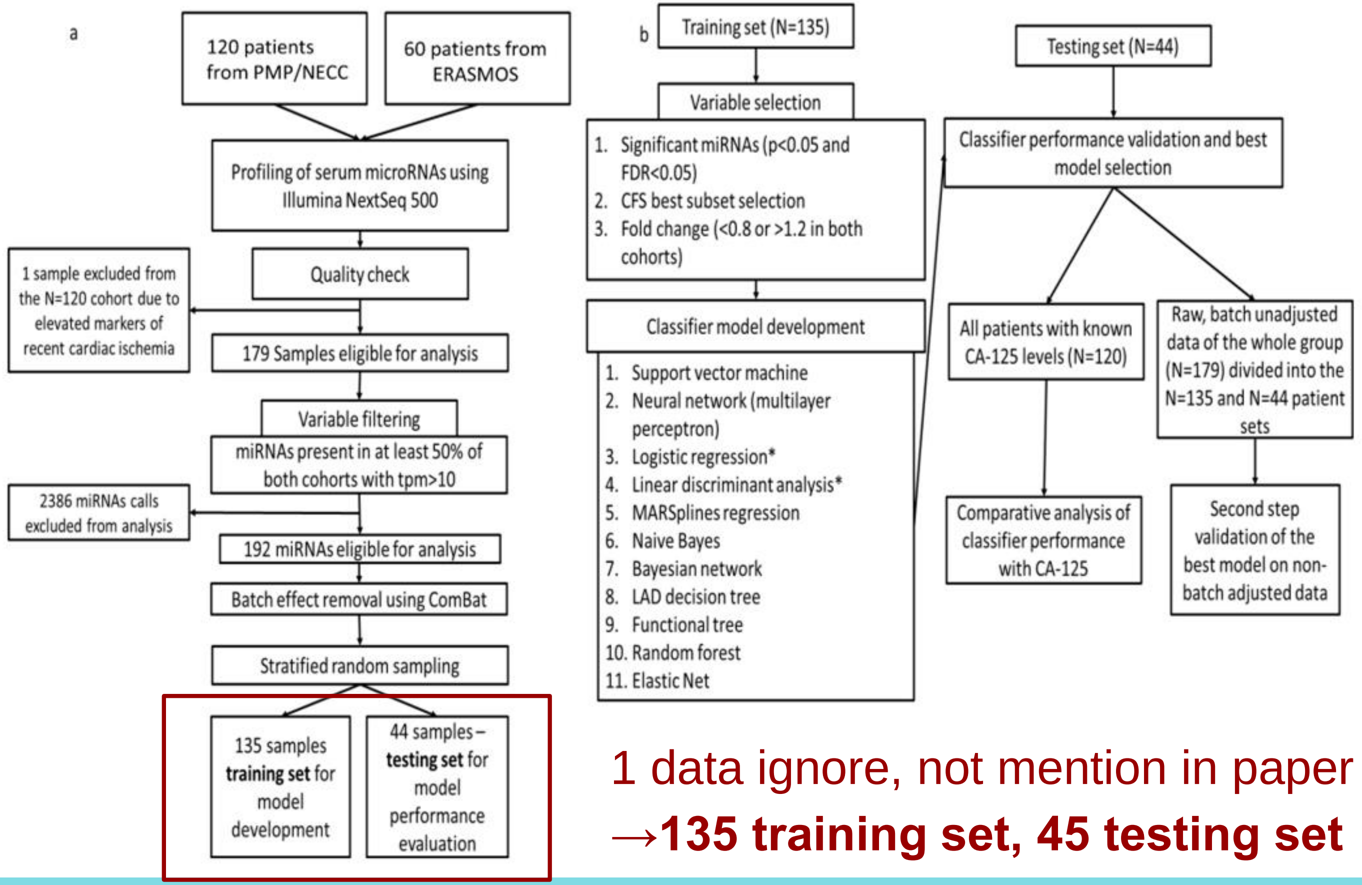


Table 3. miRNA variables used in model building identified through univariate testing

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

Split Training and Testing Set


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)
Raw, batch unadjusted data of the whole group (N=179) divided into the N=135 and N=44 patient sets
Comparative analysis of classifier performance with CA-125
Second step validation of the best model on non-batch adjusted data

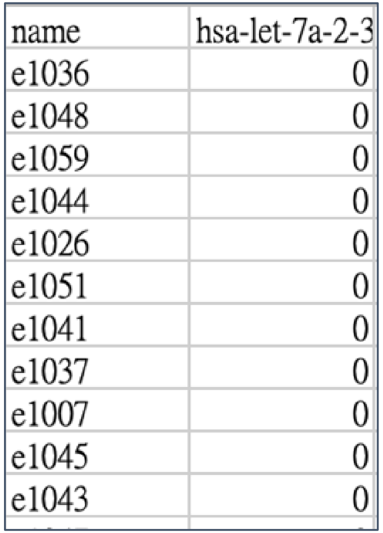
1 data ignore, not mention in paper

→**135 training set, 45 testing set**

Models

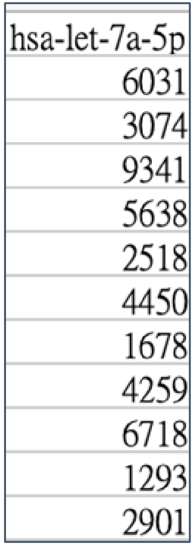
Input & Output

- Input: A person with selected miRNA features.
- Output: 1 for cancer and 0 for benign/borderline/control.

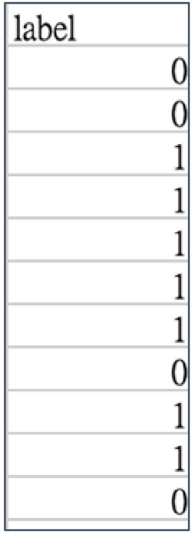


name	hsa-let-7a-2-3
e1036	0
e1048	0
e1059	0
e1044	0
e1026	0
e1051	0
e1041	0
e1037	0
e1007	0
e1045	0
e1043	0

...



hsa-let-7a-5p
6031
3074
9341
5638
2518
4450
1678
4259
6718
1293
2901



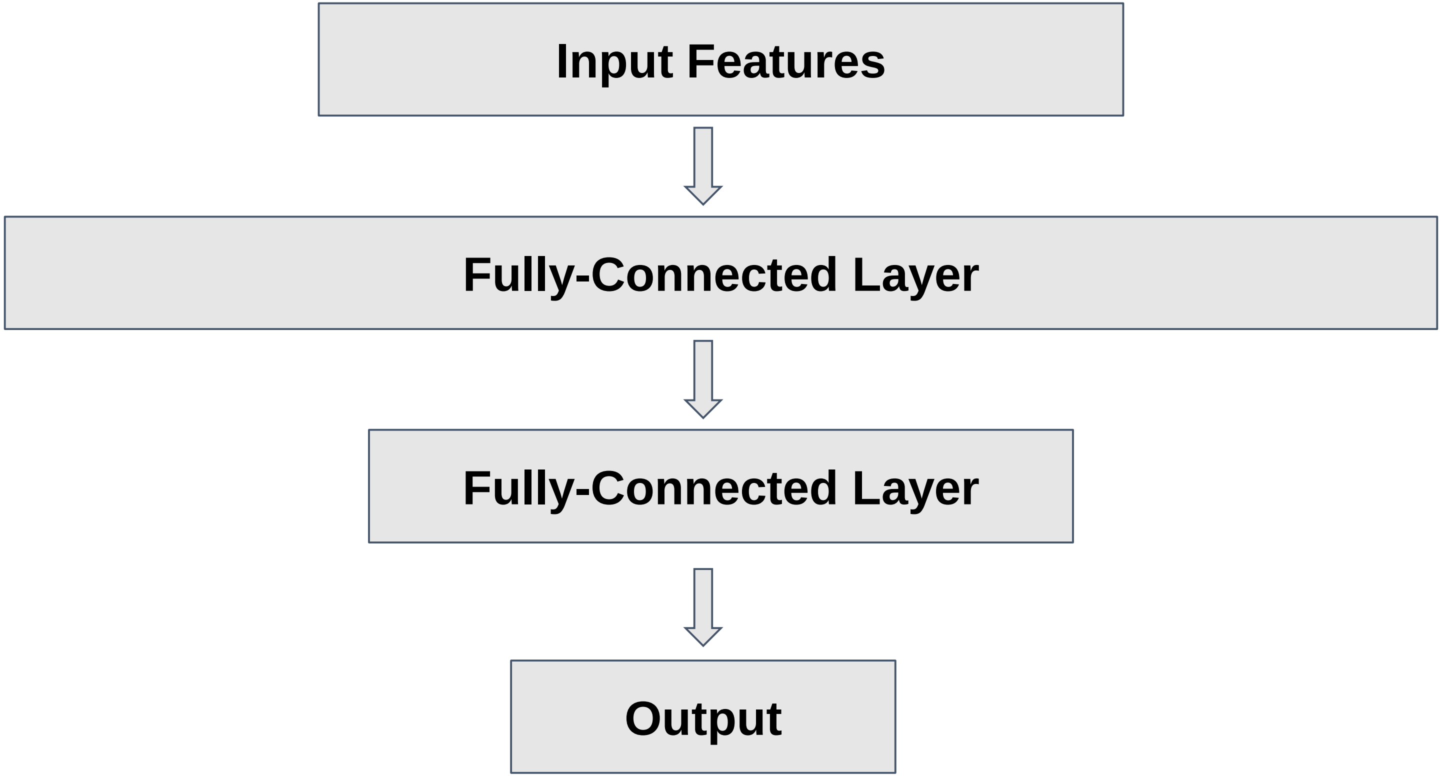
label
0
0
1
1
1
1
1
0
1
1
0

Reproduced Models

1. We reproduced several models in **scikit-learn.**
    a. Linear discriminant analysis (LDA)
    b. Logistic Regression (LR)
    c. Support vector machine (SVM)
    d. Random forest
    e. Elastic net

1. We also reproduced the NN model using **Keras**.

Neural network



Input Features

Fully-Connected Layer

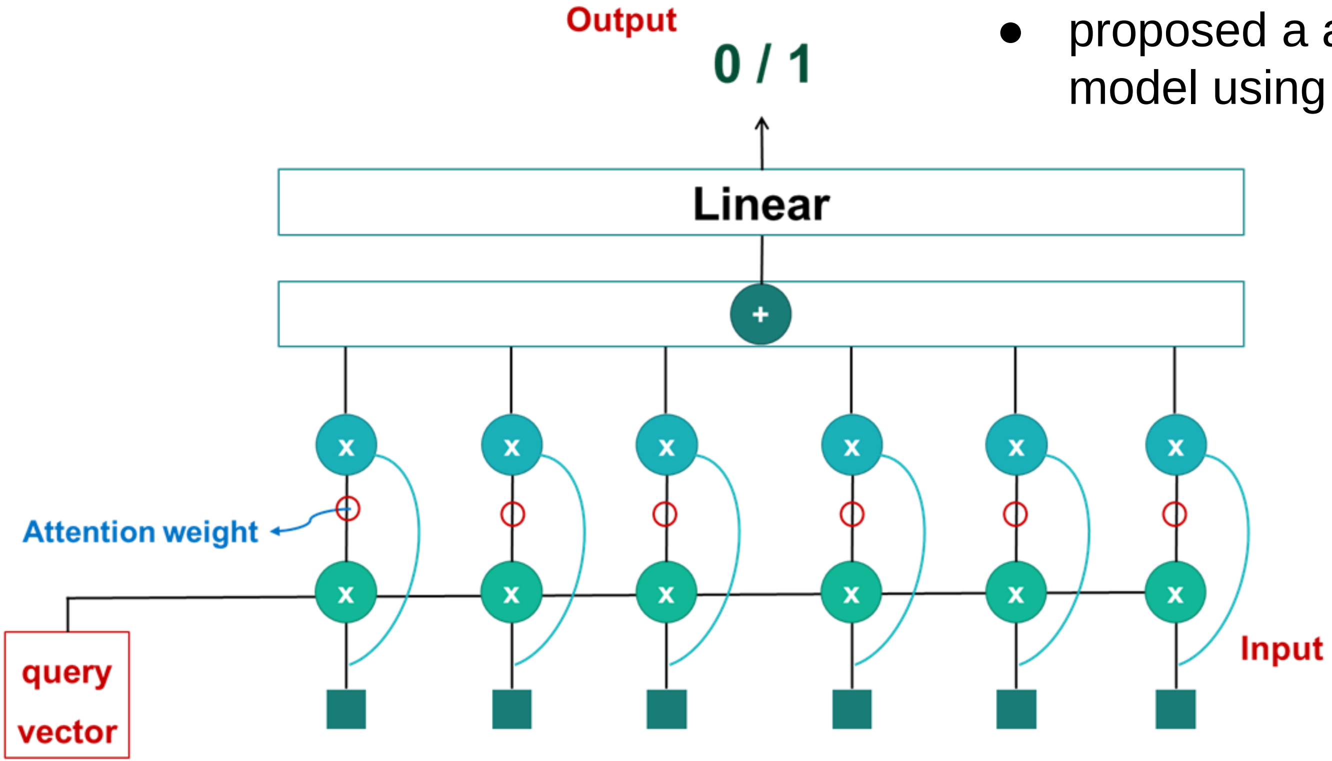
Fully-Connected Layer

Output

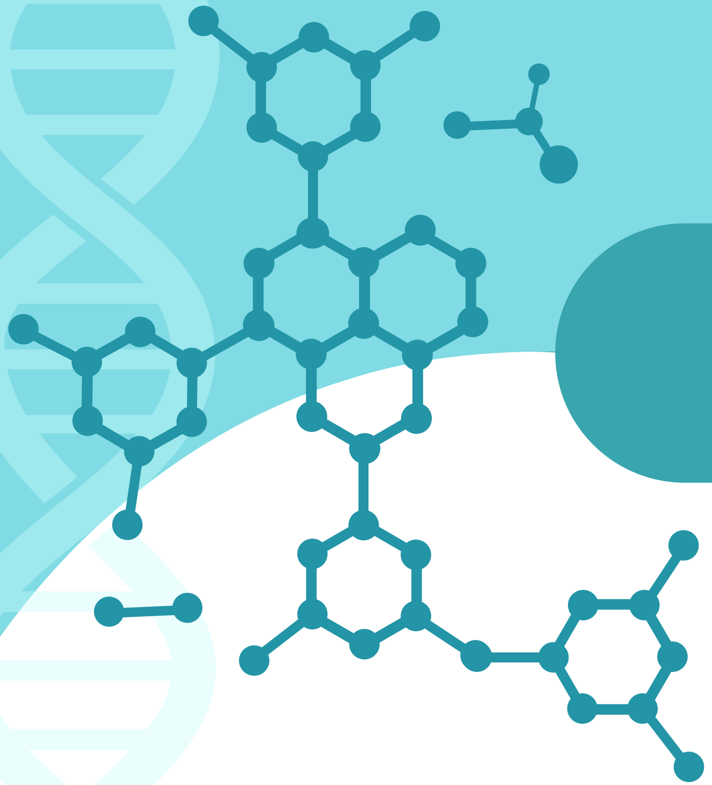
Another Model

Attention

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.



- proposed a attention-baed model using **PyTorch**



Results

Linear discriminant analysis

Paper

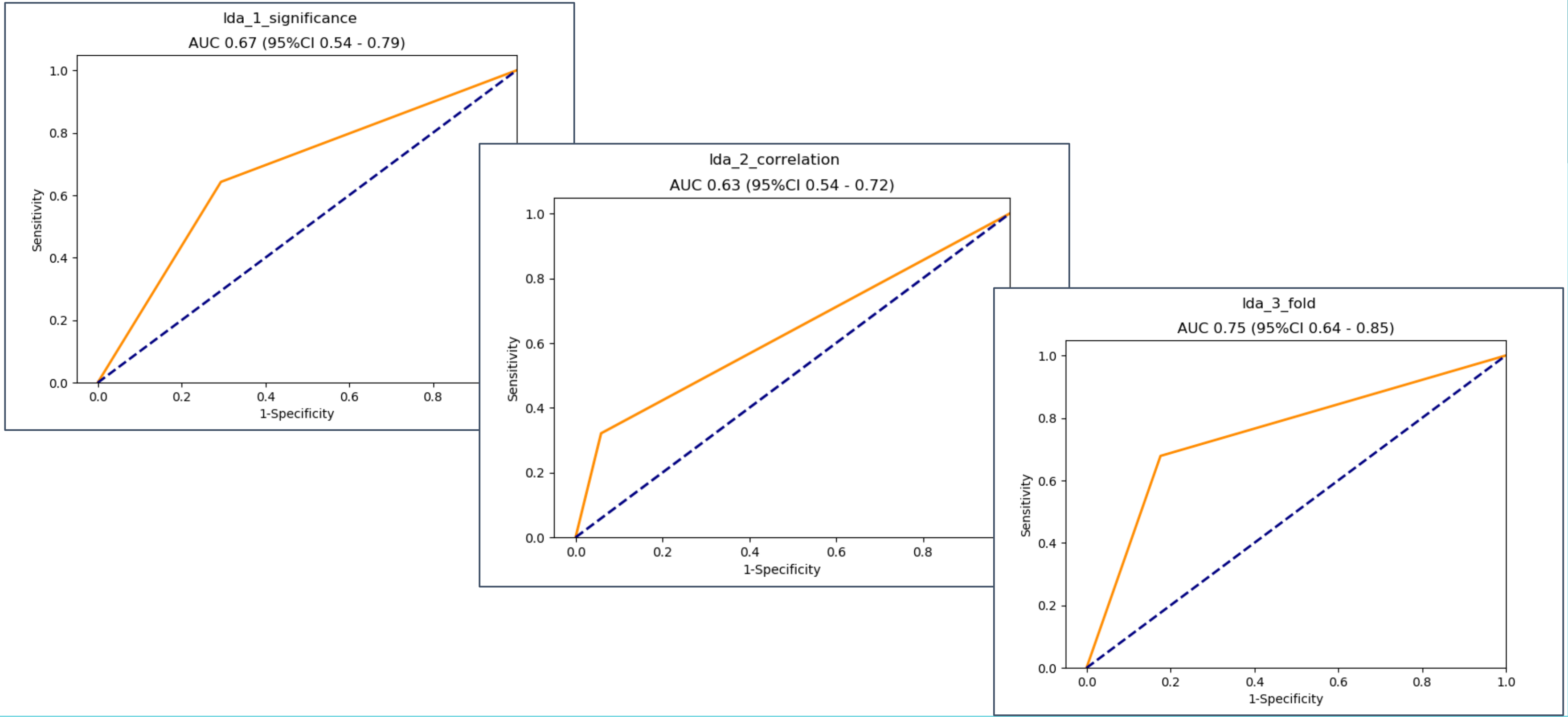
	Significance	Correlation	Fold
AUC (95% CI)	0.80 (0.66–0.93)	0.76 (0.62–0.90)	0.78 (0.64–0.92)

Ours

	Significance	Correlation	Fold
AUC (95% CI)	0.67 (0.54-0.79)	0.63 (0.54-0.72)	0.75 (0.64-0.85)

Linear discriminant analysis

Ours


lda_1_significance

AUC 0.67 (95%CI 0.54 - 0.79)

lda_2_correlation

AUC 0.63 (95%CI 0.54 - 0.72)

lda_3_fold

AUC 0.75 (95%CI 0.64 - 0.85)

Logistic regression

Paper

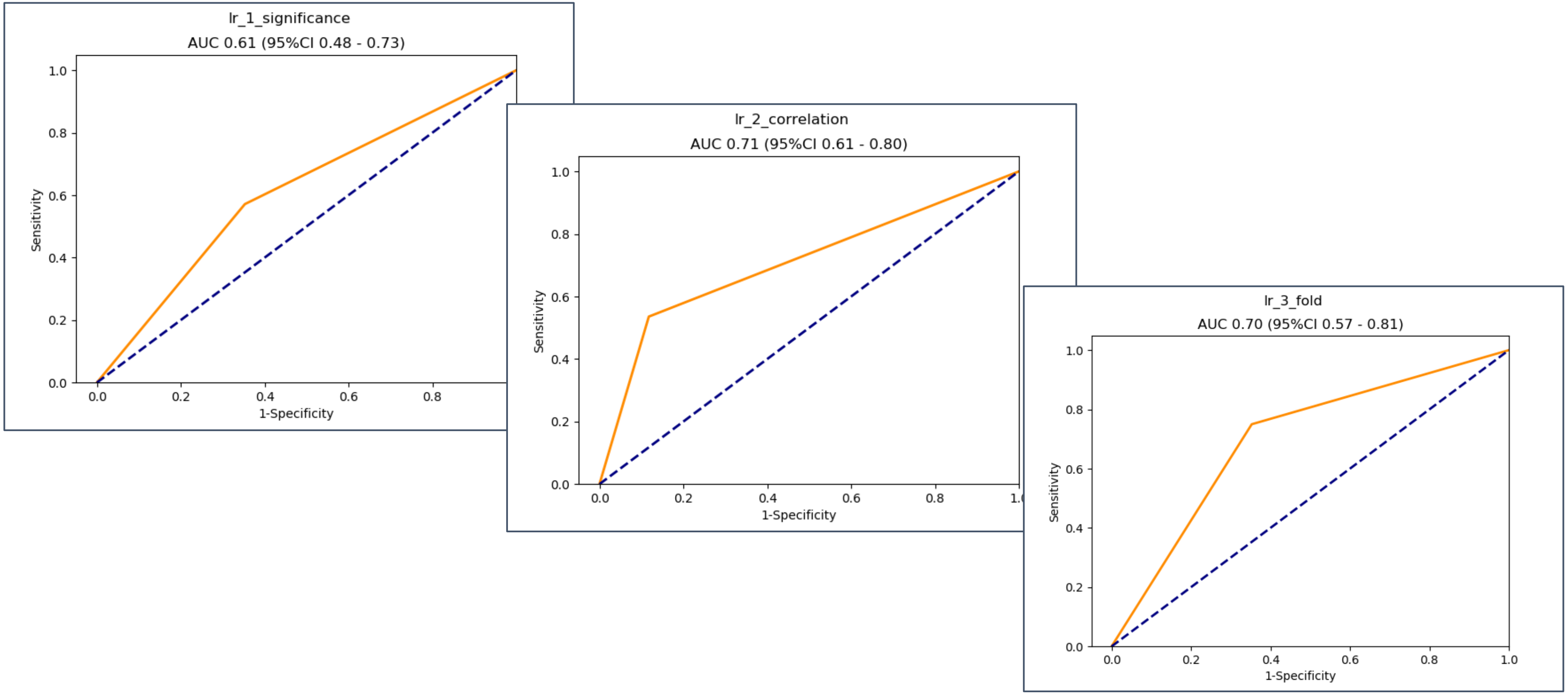
	Significance	Correlation	Fold
AUC (95% CI)	0.81 (0.68–0.94)	0.75 (0.61–0.90)	0.82 (0.70–0.94)

Ours

	Significance	Correlation	Fold
AUC (95% CI)	0.61 (0.48-0.73)	0.71 (0.61-0.80)	0.70 (0.57-0.81)

Logistic regression

Ours



Support vector machine

Paper

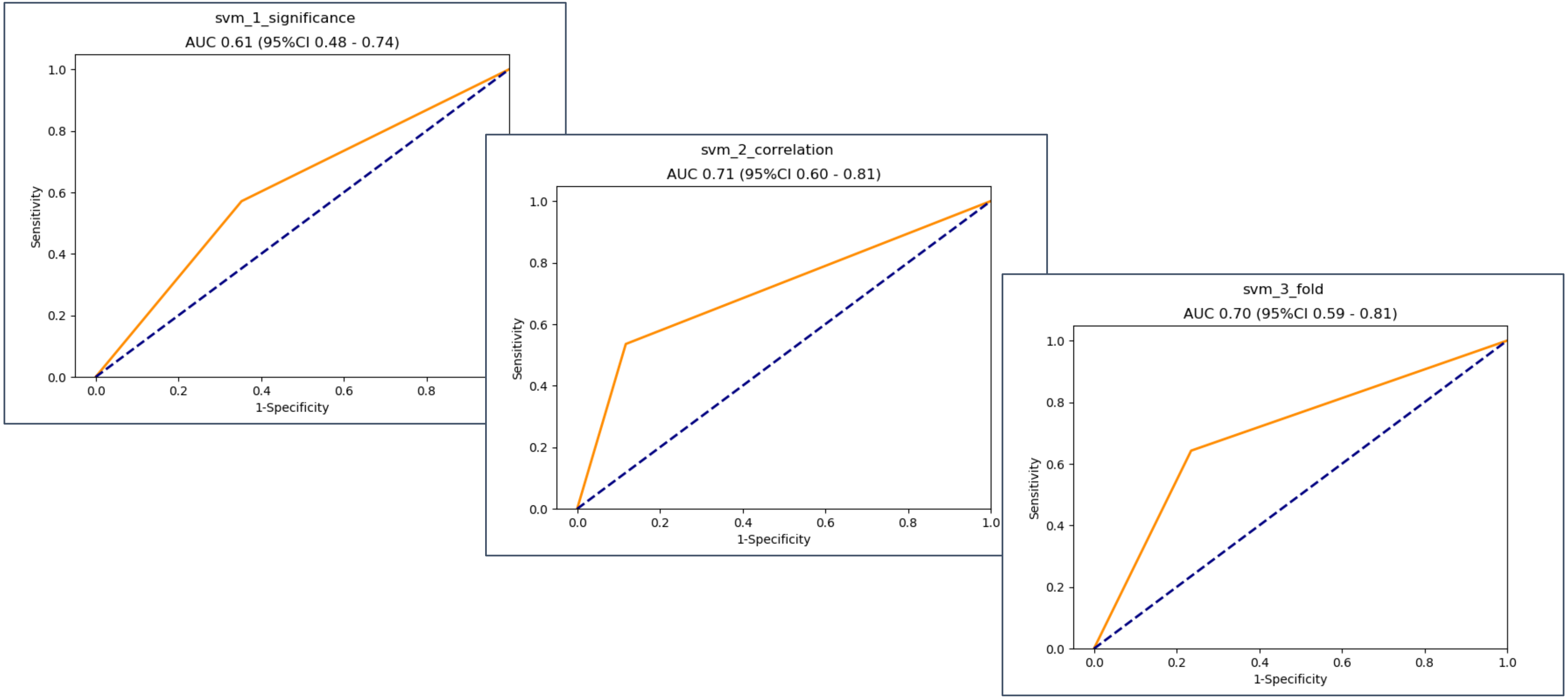
	Significance	Correlation	Fold
AUC (95% CI)	0.77 (0.63–0.91)	0.73 (0.58–0.87)	0.77 (0.63–0.91)

Ours

	Significance	Correlation	Fold
AUC (95% CI)	0.61 (0.48-0.74)	0.71 (0.60-81)	0.70 (0.59-0.81)

Support vector machine

Ours



svm_1_significance

AUC 0.61 (95%CI 0.48 - 0.74)

svm_2_correlation

AUC 0.71 (95%CI 0.60 - 0.81)

svm_3_fold

AUC 0.70 (95%CI 0.59 - 0.81)

Random forest

Paper

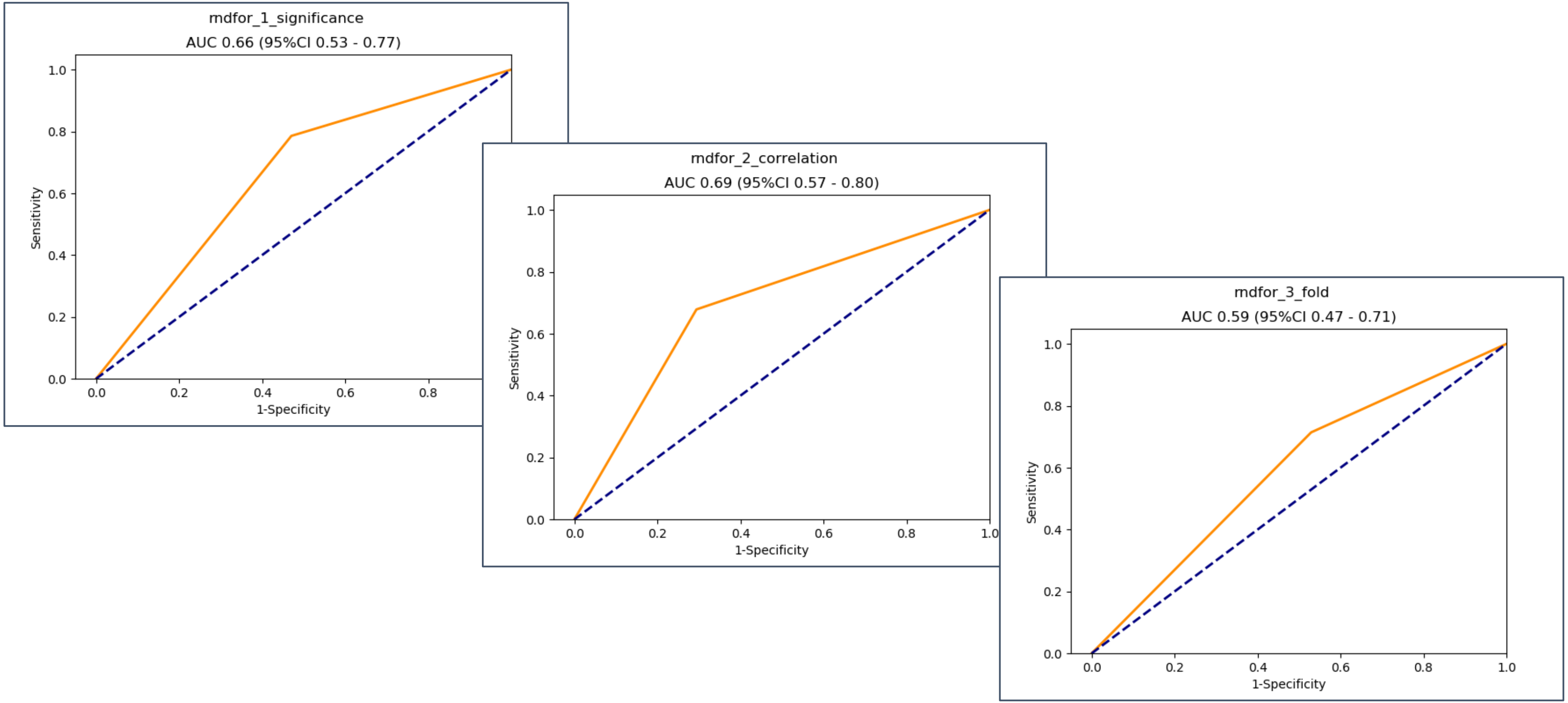
	Significance	Correlation	Fold
AUC (95% CI)	0.78 (0.64–0.91)	0.71 (0.56–0.86)	0.76 (0.62–0.90)

Ours

	Significance	Correlation	Fold
AUC (95% CI)	0.66 (0.53-0.77)	0.69 (0.57-0.80)	0.59 (0.47-0.71)

Random forest

Ours



Elastic net

Paper

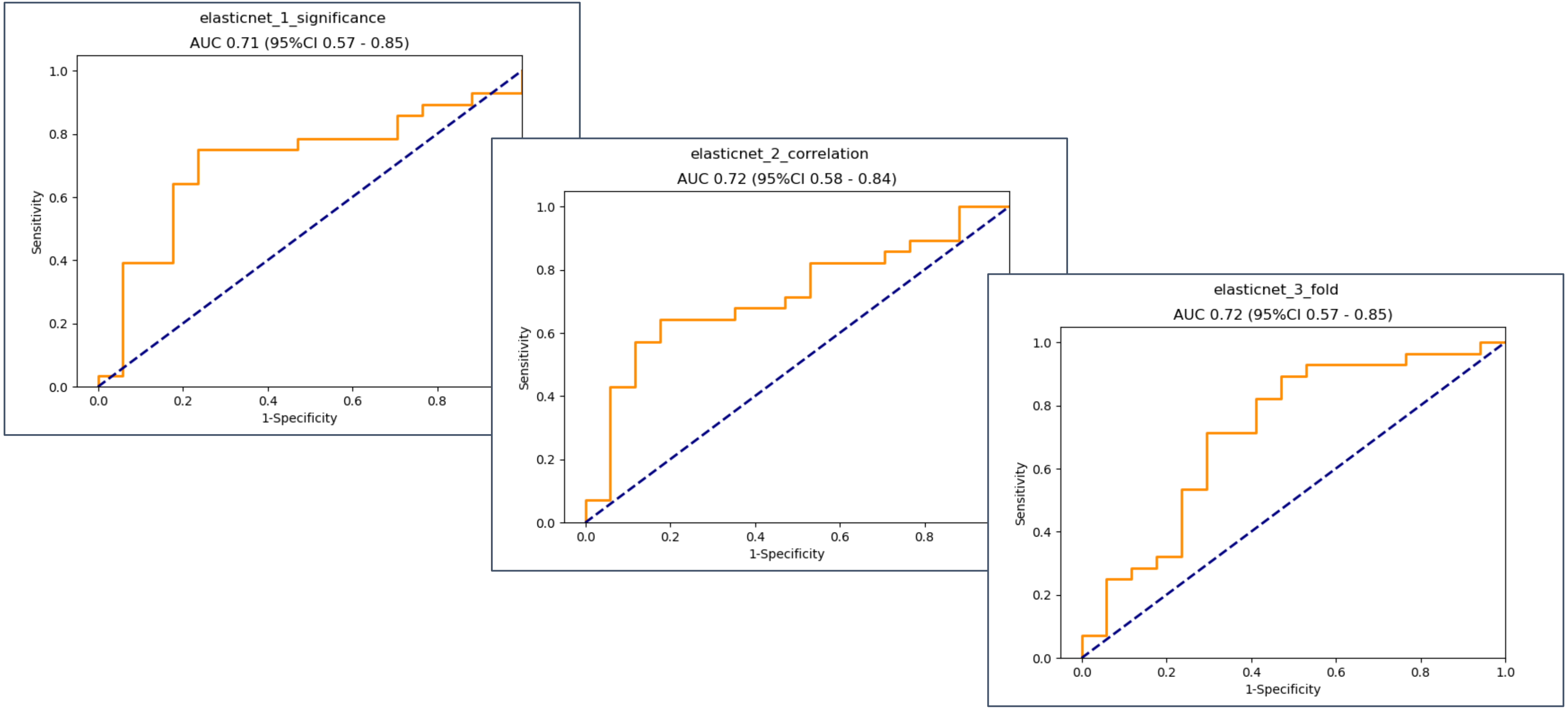
	Significance	Correlation	Fold
AUC (95% CI)	0.80 (0.67–0.93)	0.76 (0.62–0.90)	0.79 (0.66–0.92)

Ours

	Significance	Correlation	Fold
AUC (95% CI)	0.71 (0.57-0.85)	0.72 (0.58-0.84)	0.72 (0.57-0.85)

Elastic net

Ours



Neural network (3-layer)

Paper

	Significance	Correlation	Fold
AUC (95% CI)	0.84 (0.72–0.96)	0.75 (0.60–0.89)	0.90 (0.81–0.99)

Ours

	Significance	Correlation	Fold
AUC (95% CI)	0.65 (0.58-0.72)	0.57 (0.45-0.69)	0.63 (0.51-0.76)

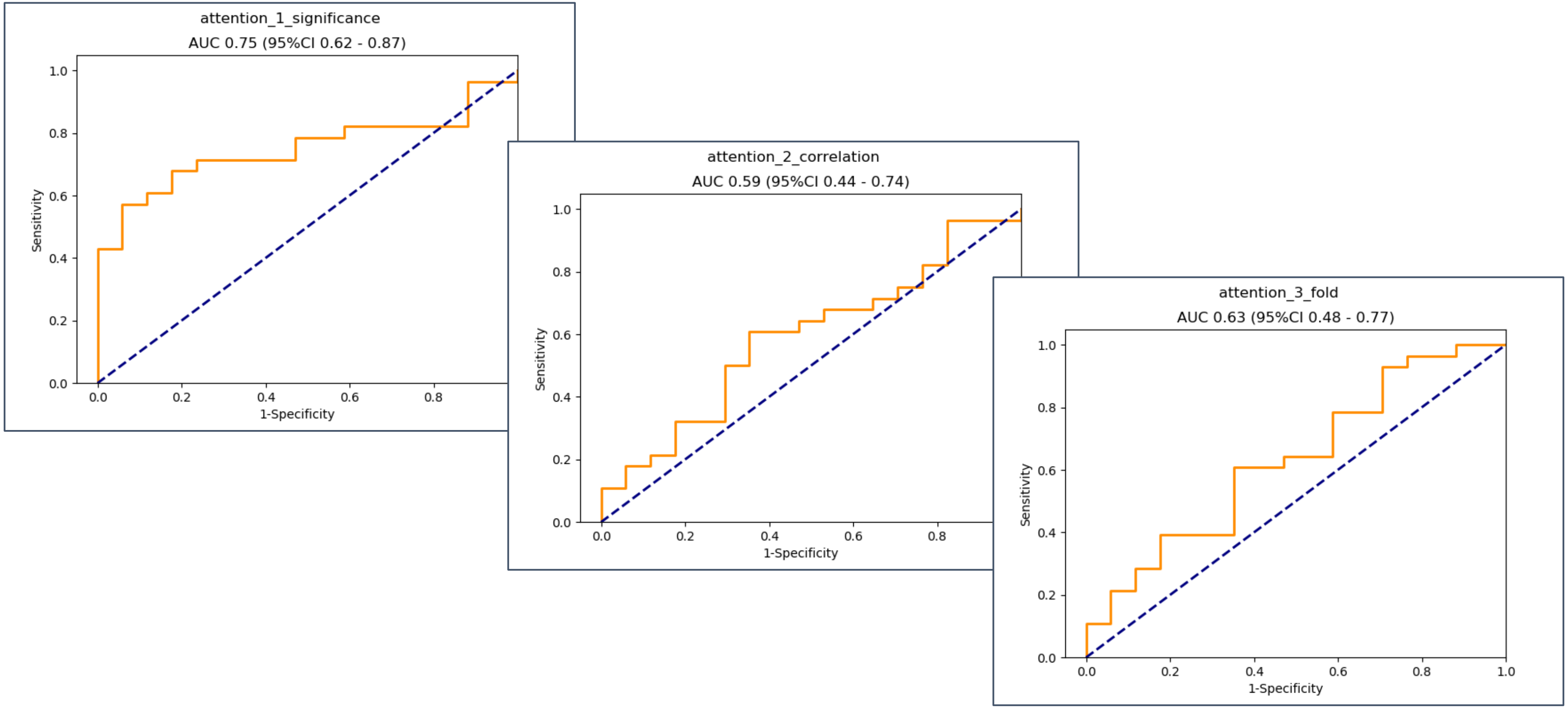
Attention

ours

	Significance	Correlation	Fold
AUC (95% CI)	**0.75 (0.62-0.87)**	0.59 (0.44-0.74)	0.63 (0.48-0.77)

Attention

Ours



attention_1_significance
AUC 0.75 (95%CI 0.62 - 0.87)

attention_2_correlation
AUC 0.59 (95%CI 0.44 - 0.74)

attention_3_fold
AUC 0.63 (95%CI 0.48 - 0.77)

Conclusion

1. The original authors uses STATISTICA and Weka to produce their models. Since these programs either required commercial licenses or is hare to use, we chose to use python to reproduce the results. However, we are not able to reproduce the results since STATISTICA and Weka provides some fine-tuning that we don't know.

2. We proposed a attention-baed model using PyTorch, and gained a good enough results. However, the results are still worse than the best models proposed by the original authors, maybe because of the lack of data complexity, witch cause our model easily to overfit the training data.

Others

Corporate

(1) 資料下載及處理：段寶鈞、陳先灝、盧佳妤

(2) 模型重現：

- (a) 挑特徵值：陳先灝
- (b) 模型建構及結果：
  - LDA: 盧佳妤
  - LR: 段寶鈞
  - SVM: 盧佳妤
  - Random forest: 盧佳妤
  - Elastic net: 盧佳妤
  - NN: 陳先灝

(3) 其他模型實作(attention)：盧佳妤

References

* Paper

- Elias, K. M. et al. Diagnostic potential for a serum miRNA neural network for detection of ovarian cancer. Elife 6, (2017)
- Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

* Packages we use

- numpy (matrix utilities)
- scikit-learn (statistics routines)
- pytorch (attention model framework)
- keras (NN model framework)
- matplotlib (figure drawing)
- absl-py (flag management)
- coloredlogs (beautiful logging)

Thank You