TAN TAO UNIVERSITY
FROM KNOWLEDGE TO THE STARS

Serum Levels of 2 miRNA- A High Accuracy Signature For Early Prostate Cancer Detection Using Random Forest

Thanh-Giang Tan Nguyen
Tan Tao University, School of Biotechnology
Instructor:
Minh-Nam Nguyen, PhD
School of Medicine, Vietnam National University, Ho Chi Minh city

*Outline:

1. Introduction

- Current methods for prostate cancer diagnosis: **<u>The challenges and opportunities.</u>**

2. Material and methods

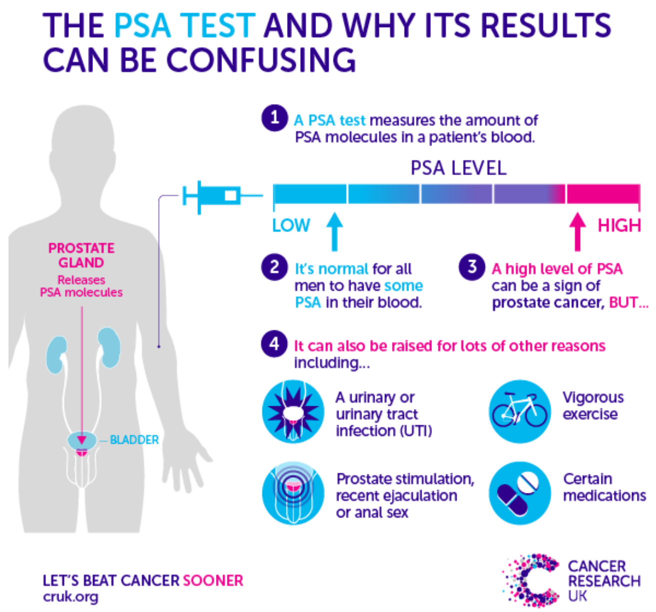
- **<u>The 1231 miRNA samples</u>** from public datasets GSE112264 and GSE11386 (https://www.ncbi.nlm.nih.gov/geo/)

3. Results

- Differential gene expression between prostate cancer and non prostate cancer control.
- **<u>Benchmarking machine learning techniques for high accuracy diagnosis models</u>**

4. Conclusions

- Developed successfully models **<u>state of the art using only 2 miRNA gene expressions.</u>**



THE PSA TEST AND WHY ITS RESULTS CAN BE CONFUSING

1 A PSA test measures the amount of PSA molecules in a patient's blood.

PSA LEVEL

LOW

HIGH

PROSTATE GLAND Releases PSA molecules

BLADDER

2 It's normal for all men to have some PSA in their blood.

3 A high level of PSA can be a sign of prostate cancer, BUT...

4 It can also be raised for lots of other reasons including...

A urinary or urinary tract infection (UTI)

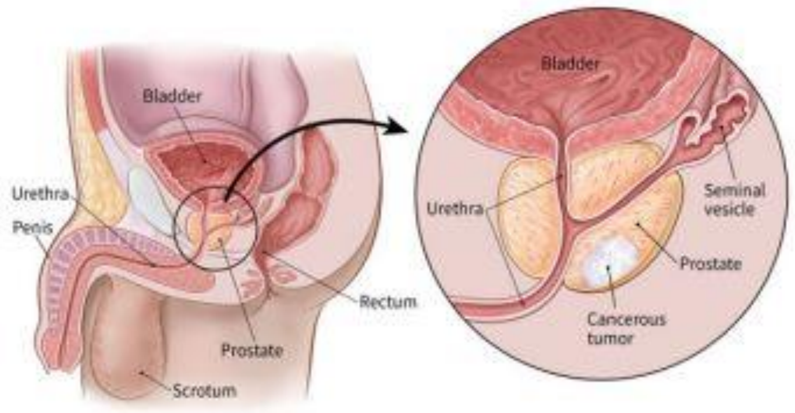
Vigorous exercise

Prostate stimulation, recent ejaculation or anal sex

Certain medications

LET'S BEAT CANCER SOONER cruk.org

CANCER RESEARCH UK



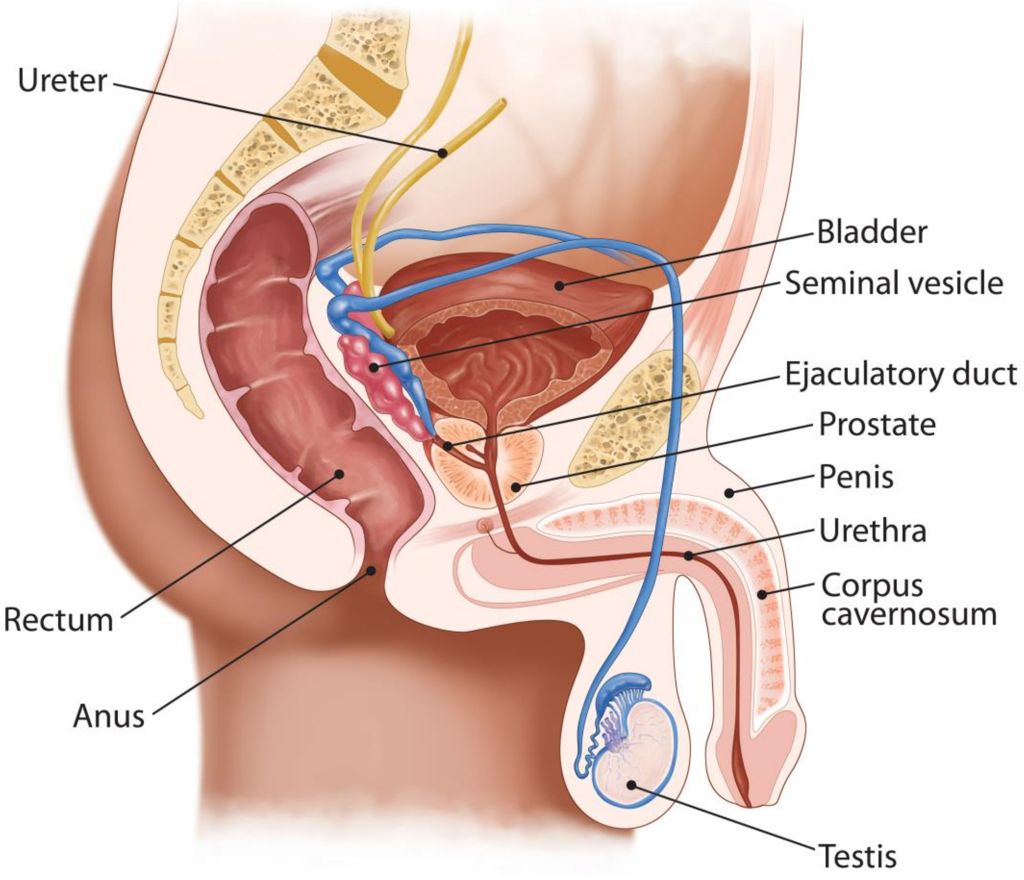
https://www.medicinenet.com/prostate_specific_antigen/article.htm

1. Introduction:

Prostate disease

- The prostate is a walnut-sized gland in the reproductive system of men.
- Prostate Cancer is **the most common cancer** and the **second leading cause of cancer death in men**.
- Prostate cancer: normal cells in the prostate begin to change and grow uncontrollably.

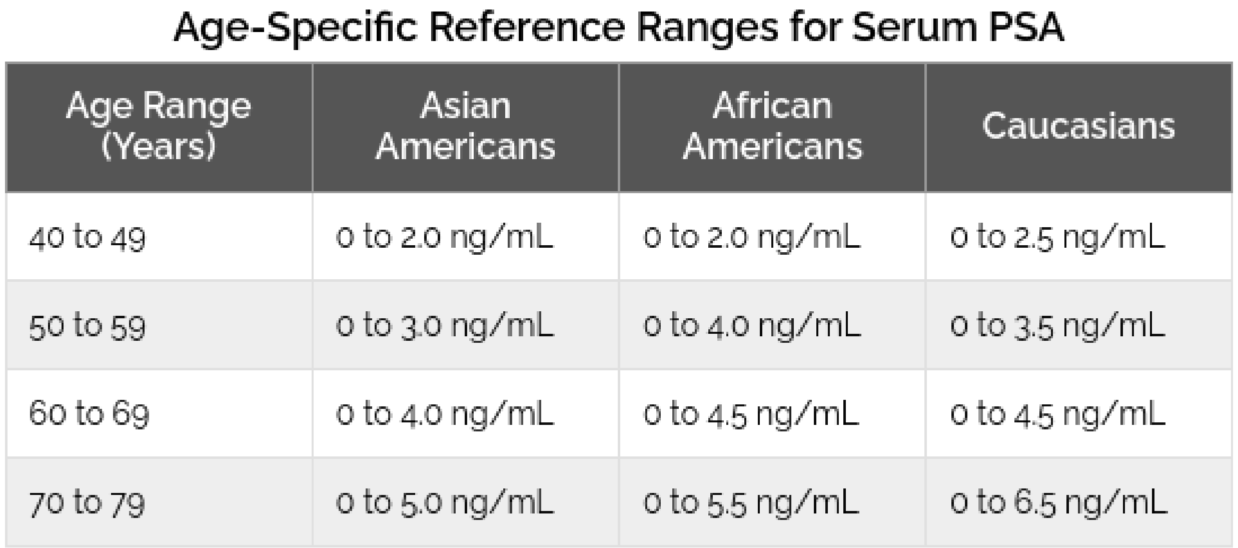
- 1 in 41 men will die of this disease.
- The incidence of prostate cancer increases with age or in men who have a history of this disease in their family.
- **Nearly 100% survivable if caught early.**



https://www.prostateconditions.org/about-prostate-conditions/prostate-cancer?gclid=Cj0KCQjw_4-SBhCgARIsAAlegrV9fYBXLK04hagv_dvUd23NEzVstDDuP-CZnZWuW0tnwShDvCeoF8oaApm-EALw_wcB

1. Introduction:

Prostate cancer, current methods for detection

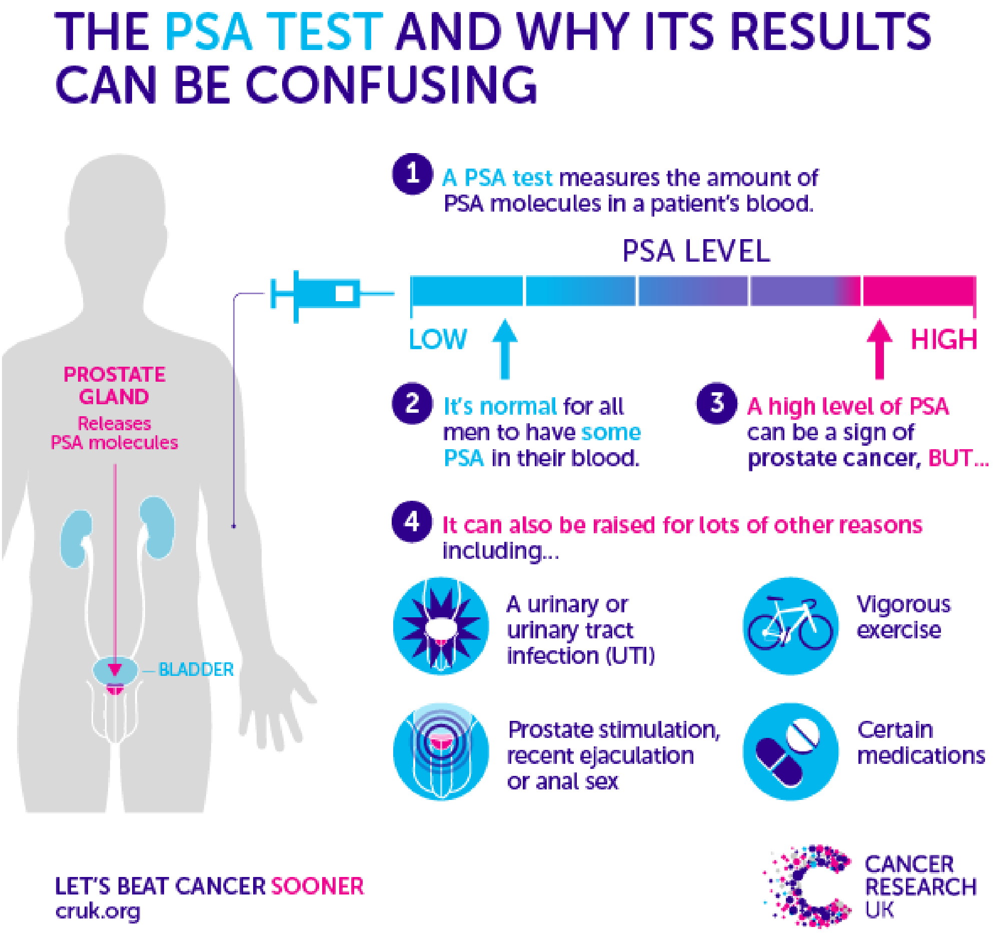


Age-Specific Reference Ranges for Serum PSA

Age Range (Years)	Asian Americans	African Americans	Caucasians
40 to 49	0 to 2.0 ng/mL	0 to 2.0 ng/mL	0 to 2.5 ng/mL
50 to 59	0 to 3.0 ng/mL	0 to 4.0 ng/mL	0 to 3.5 ng/mL
60 to 69	0 to 4.0 ng/mL	0 to 4.5 ng/mL	0 to 4.5 ng/mL
70 to 79	0 to 5.0 ng/mL	0 to 5.5 ng/mL	0 to 6.5 ng/mL

- PSA test is most popular
- The Prostate-specific antigen (PSA) test is **a blood test to help detect prostate cancer**.
- 70% to 80% of men with an elevated PSA who have a biopsy do not have cancer

https://www.health.harvard.edu/blog/is-psa-reliable-20110327214



THE PSA TEST AND WHY ITS RESULTS CAN BE CONFUSING

1 A PSA test measures the amount of PSA molecules in a patient's blood.

2 It's normal for all men to have some PSA in their blood.

3 A high level of PSA can be a sign of prostate cancer, BUT...

4 It can also be raised for lots of other reasons including...

A urinary or urinary tract infection (UTI)

Vigorous exercise

Prostate stimulation, recent ejaculation or anal sex

Certain medications

LET'S BEAT CANCER SOONER
cruk.org

1. Introduction:

However,...

The limitations of PSA testing:

- **Factors that increase PSA:** prostate enlargement, prostate inflammation, infection, or increase with age.
- **Factors that decrease PSA**. Dosages of some medications used for urinary conditions can lower PSA levels. Obesity can also lower PSA levels => False results

Risks:

- **Biopsy problems** include bleeding and infection.
- **Psychological effects** on the person diagnosed due to false positives.

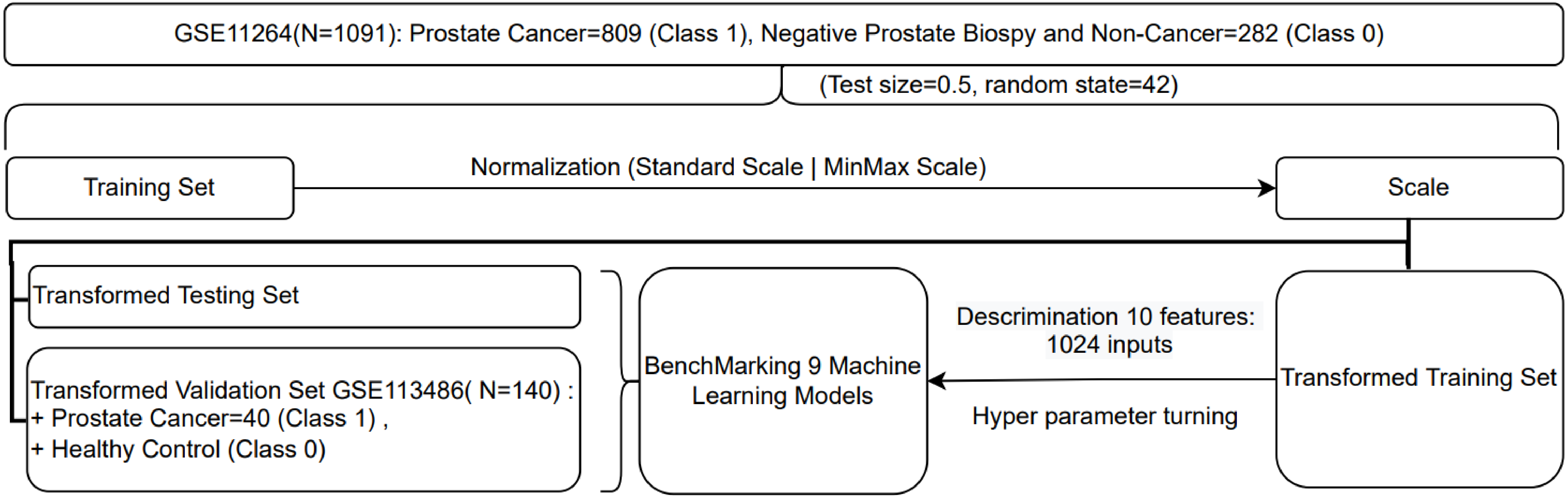
=> The method uses miRNA with **higher accuracy,** which is **noninvasive**

https://www.mayoclinic.org/tests-procedures/psa-test/about/pac-20384731#:~:text=The%20limitations%20of%20PSA%20testing,levels%20normally%20increase%20with%20age.

2. Materials and methods

Data: Using public datasets with miRNA **gene expression array processed data**

Workflow For Early Detection Prostate Cancer Using 2565 miRNAs



GSE11264(N=1091): Prostate Cancer=809 (Class 1), Negative Prostate Biospy and Non-Cancer=282 (Class 0)

(Test size=0.5, random state=42)

Training Set

Normalization (Standard Scale | MinMax Scale)

Scale

Transformed Testing Set

Transformed Validation Set GSE113486(N=140) :
+ Prostate Cancer=40 (Class 1) ,
+ Healthy Control (Class 0)

BenchMarking 9 Machine Learning Models

Descrimination 10 features: 1024 inputs

Hyper parameter turning

Transformed Training Set

2. Materials and methods

Statistical analysis and visualization:

R version 4.0.1 packages:
Data visualization: ggpubr
Python version 3.9.5:
Data preprocess and model predictions

Webserver:
+ miRNA target prediction:
https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/index.php
+Gene Ontology enrichment analysis:
http://bioinformatics.sdstate.edu/go/

Notes: All the required environments, packages and scripts are available at:
https://github.com/nttg8100/Highly-Accurate-Early-Detection-using-miRNA-gene-expression-panel-in-Prostate-Cancer.git

3.Results

Features selection for prostate cancer diagnosis model

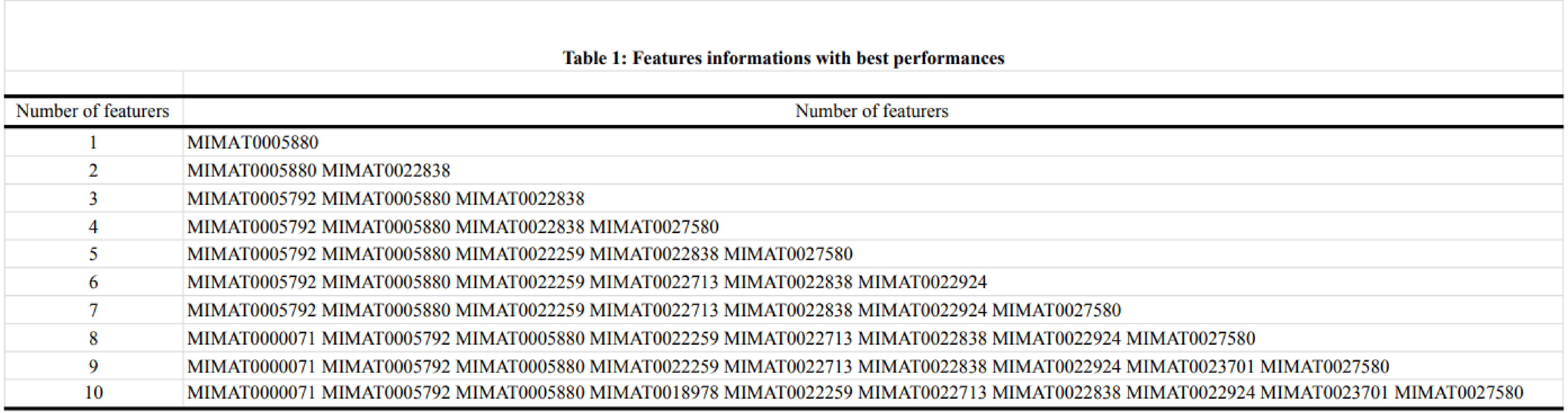


Table 1: Features informations with best performances

Number of featurers	Number of featurers
1	MIMAT0005880
2	MIMAT0005880 MIMAT0022838
3	MIMAT0005792 MIMAT0005880 MIMAT0022838
4	MIMAT0005792 MIMAT0005880 MIMAT0022838 MIMAT0027580
5	MIMAT0005792 MIMAT0005880 MIMAT0022259 MIMAT0022838 MIMAT0027580
6	MIMAT0005792 MIMAT0005880 MIMAT0022259 MIMAT0022713 MIMAT0022838 MIMAT0022924
7	MIMAT0005792 MIMAT0005880 MIMAT0022259 MIMAT0022713 MIMAT0022838 MIMAT0022924 MIMAT0027580
8	MIMAT0000071 MIMAT0005792 MIMAT0005880 MIMAT0022259 MIMAT0022713 MIMAT0022838 MIMAT0022924 MIMAT0027580
9	MIMAT0000071 MIMAT0005792 MIMAT0005880 MIMAT0022259 MIMAT0022713 MIMAT0022838 MIMAT0022924 MIMAT0023701 MIMAT0027580
10	MIMAT0000071 MIMAT0005792 MIMAT0005880 MIMAT0018978 MIMAT0022259 MIMAT0022713 MIMAT0022838 MIMAT0022924 MIMAT0023701 MIMAT0027580

Note: top 10 miRNAs with highest score detected by Anova test

3.Results

Model selection for prostate cancer diagnosis model

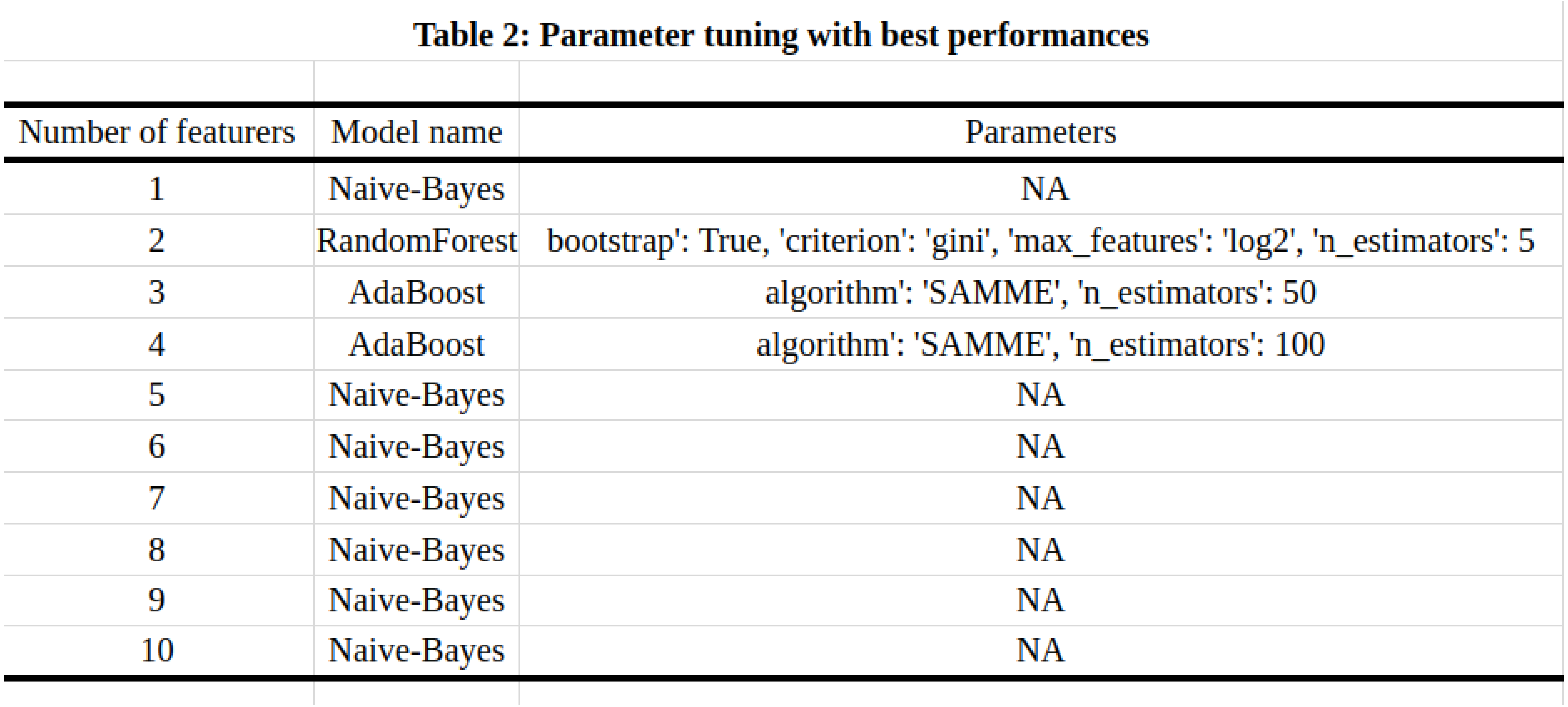


Table 2: Parameter tuning with best performances

Number of featurers	Model name	Parameters
1	Naive-Bayes	NA
2	RandomForest	bootstrap': True, 'criterion': 'gini', 'max_features': 'log2', 'n_estimators': 5
3	AdaBoost	algorithm': 'SAMME', 'n_estimators': 50
4	AdaBoost	algorithm': 'SAMME', 'n_estimators': 100
5	Naive-Bayes	NA
6	Naive-Bayes	NA
7	Naive-Bayes	NA
8	Naive-Bayes	NA
9	Naive-Bayes	NA
10	Naive-Bayes	NA

3.Results

Model performance for prostate cancer diagnosis model

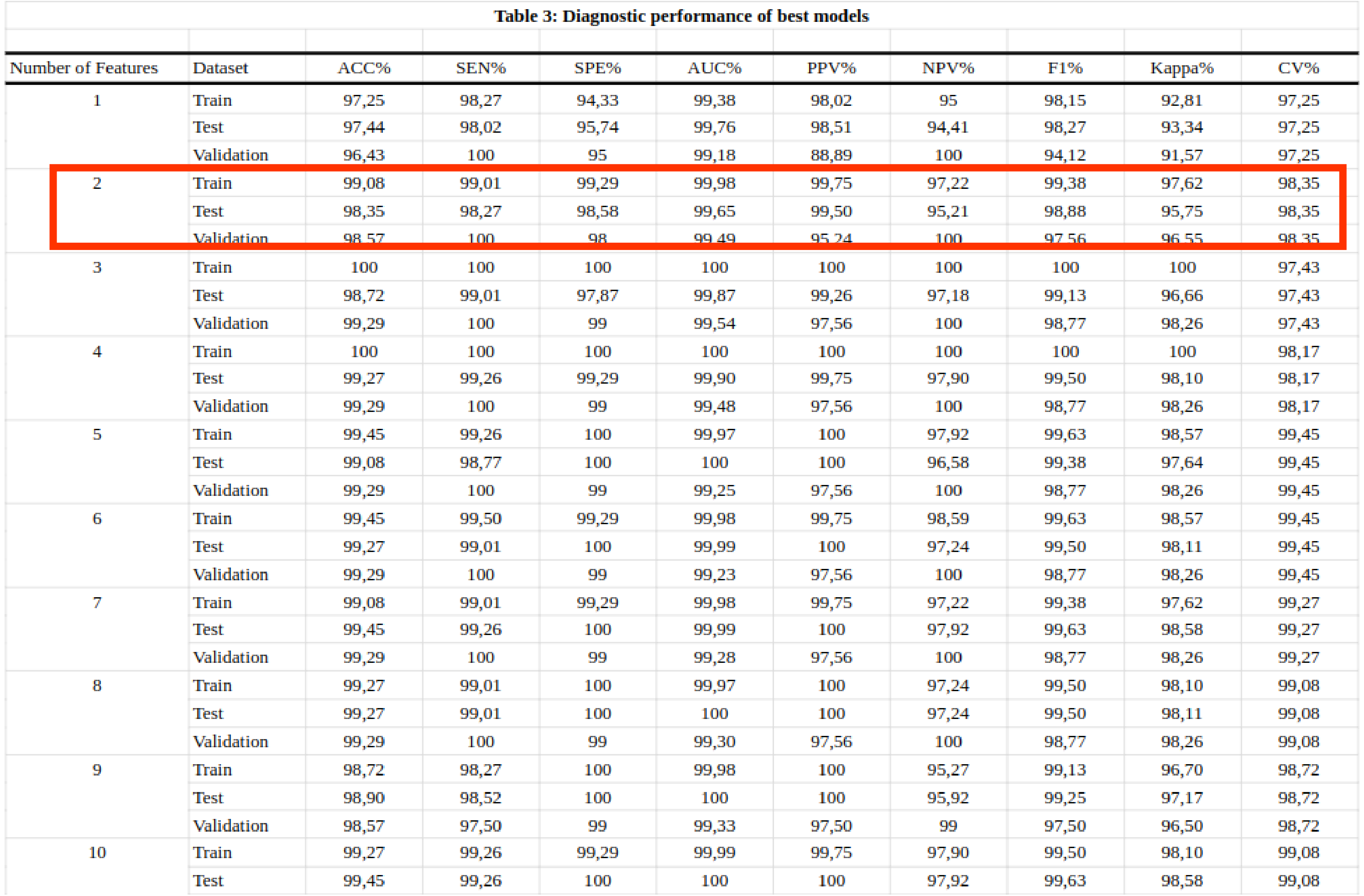
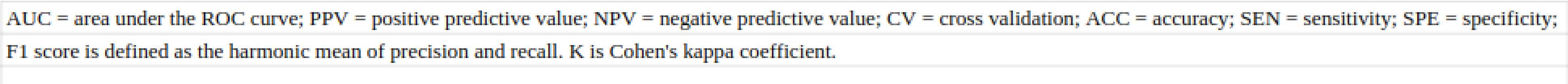


Table 3: Diagnostic performance of best models

Number of Features	Dataset	ACC%	SEN%	SPE%	AUC%	PPV%	NPV%	F1%	Kappa%	CV%
1	Train	97,25	98,27	94,33	99,38	98,02	95	98,15	92,81	97,25
	Test	97,44	98,02	95,74	99,76	98,51	94,41	98,27	93,34	97,25
	Validation	96,43	100	95	99,18	88,89	100	94,12	91,57	97,25
2	Train	99,08	99,01	99,29	99,98	99,75	97,22	99,38	97,62	98,35
	Test	98,35	98,27	98,58	99,65	99,50	95,21	98,88	95,75	98,35
	Validation	98,57	100	98	99,49	95,24	100	97,56	96,55	98,35
3	Train	100	100	100	100	100	100	100	100	97,43
	Test	98,72	99,01	97,87	99,87	99,26	97,18	99,13	96,66	97,43
	Validation	99,29	100	99	99,54	97,56	100	98,77	98,26	97,43
4	Train	100	100	100	100	100	100	100	100	98,17
	Test	99,27	99,26	99,29	99,90	99,75	97,90	99,50	98,10	98,17
	Validation	99,29	100	99	99,48	97,56	100	98,77	98,26	98,17
5	Train	99,45	99,26	100	99,97	100	97,92	99,63	98,57	99,45
	Test	99,08	98,77	100	100	100	96,58	99,38	97,64	99,45
	Validation	99,29	100	99	99,25	97,56	100	98,77	98,26	99,45
6	Train	99,45	99,50	99,29	99,98	99,75	98,59	99,63	98,57	99,45
	Test	99,27	99,01	100	99,99	100	97,24	99,50	98,11	99,45
	Validation	99,29	100	99	99,23	97,56	100	98,77	98,26	99,45
7	Train	99,08	99,01	99,29	99,98	99,75	97,22	99,38	97,62	99,27
	Test	99,45	99,26	100	99,99	100	97,92	99,63	98,58	99,27
	Validation	99,29	100	99	99,28	97,56	100	98,77	98,26	99,27
8	Train	99,27	99,01	100	99,97	100	97,24	99,50	98,10	99,08
	Test	99,27	99,01	100	100	100	97,24	99,50	98,11	99,08
	Validation	99,29	100	99	99,30	97,56	100	98,77	98,26	99,08
9	Train	98,72	98,27	100	99,98	100	95,27	99,13	96,70	98,72
	Test	98,90	98,52	100	100	100	95,92	99,25	97,17	98,72
	Validation	98,57	97,50	99	99,33	97,50	99	97,50	96,50	98,72
10	Train	99,27	99,26	99,29	99,99	99,75	97,90	99,50	98,10	99,08
	Test	99,45	99,26	100	100	100	97,92	99,63	98,58	99,08

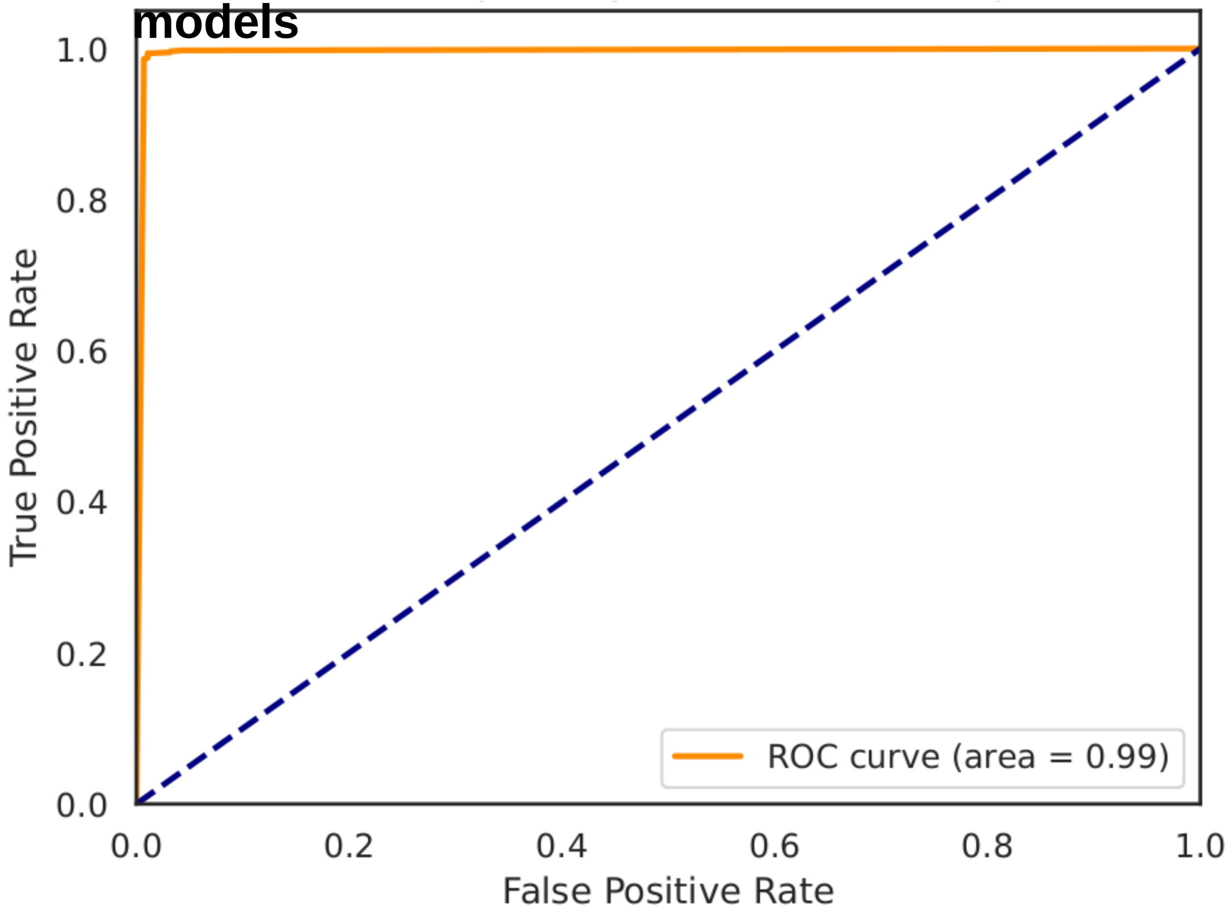


AUC = area under the ROC curve; PPV = positive predictive value; NPV = negative predictive value; CV = cross validation; ACC = accuracy; SEN = sensitivity; SPE = specificity; F1 score is defined as the harmonic mean of precision and recall. K is Cohen's kappa coefficient.

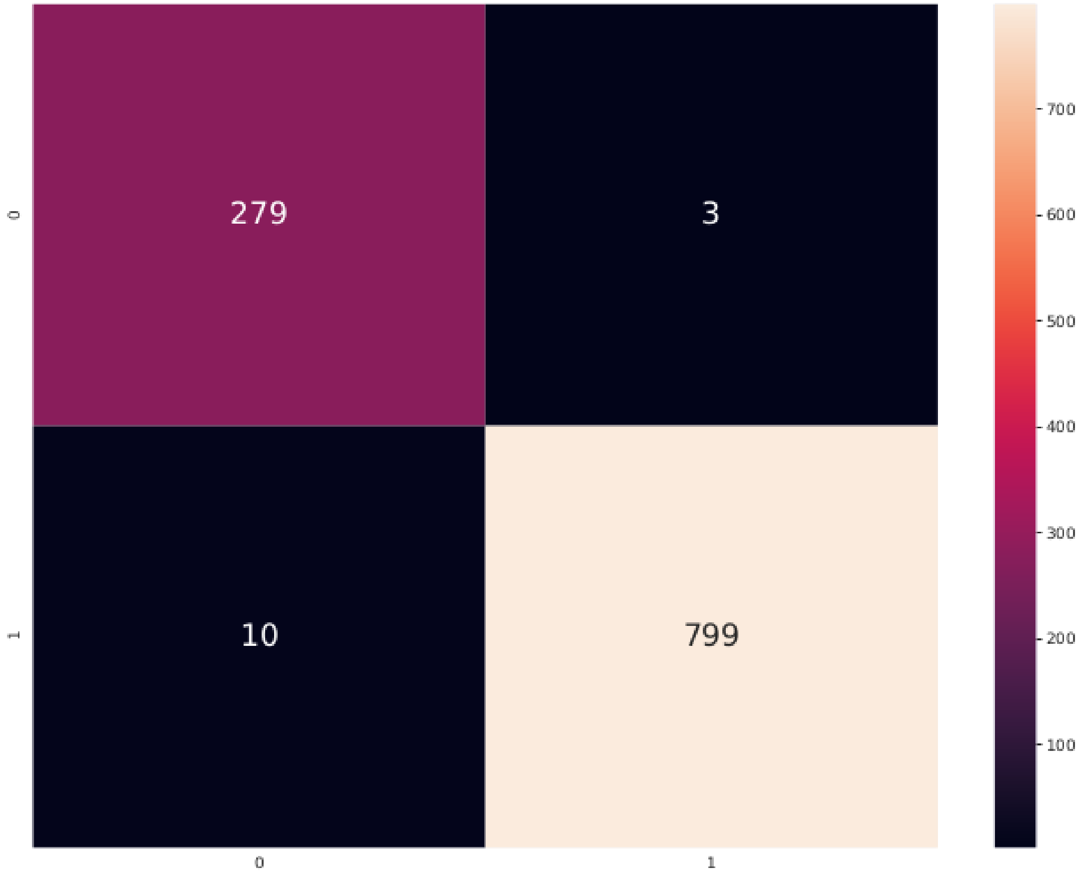
3.Results

Performance for target genes of 2 miRNAs in the final model

Receive Operating Characteristic of final models



True Class



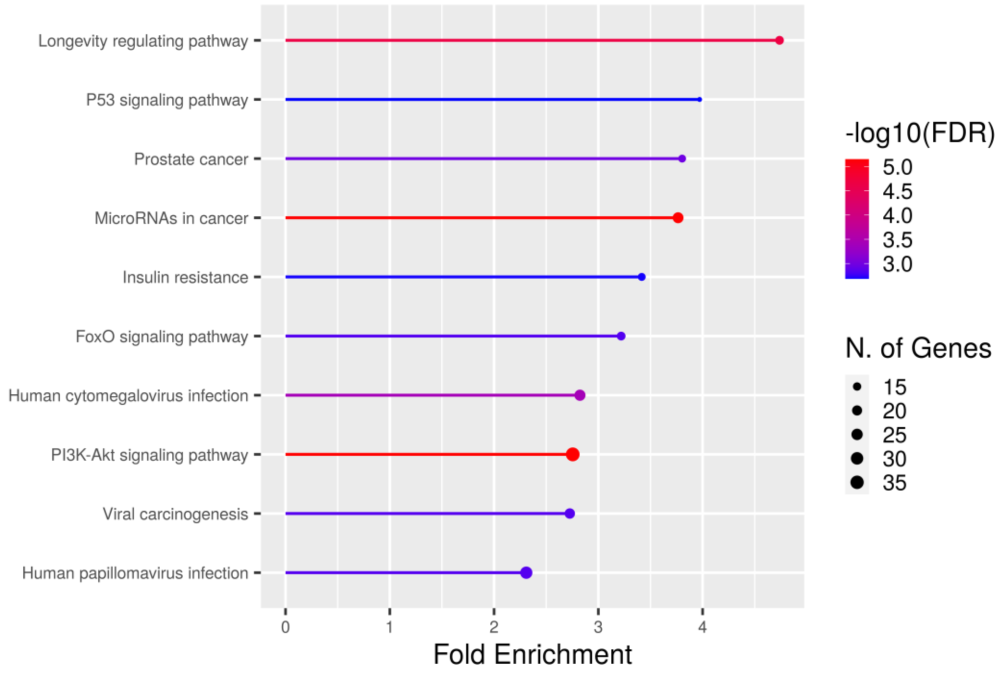
Class: 0: Negative Prostate and Non-Cancer; 1: Prostate Cancer

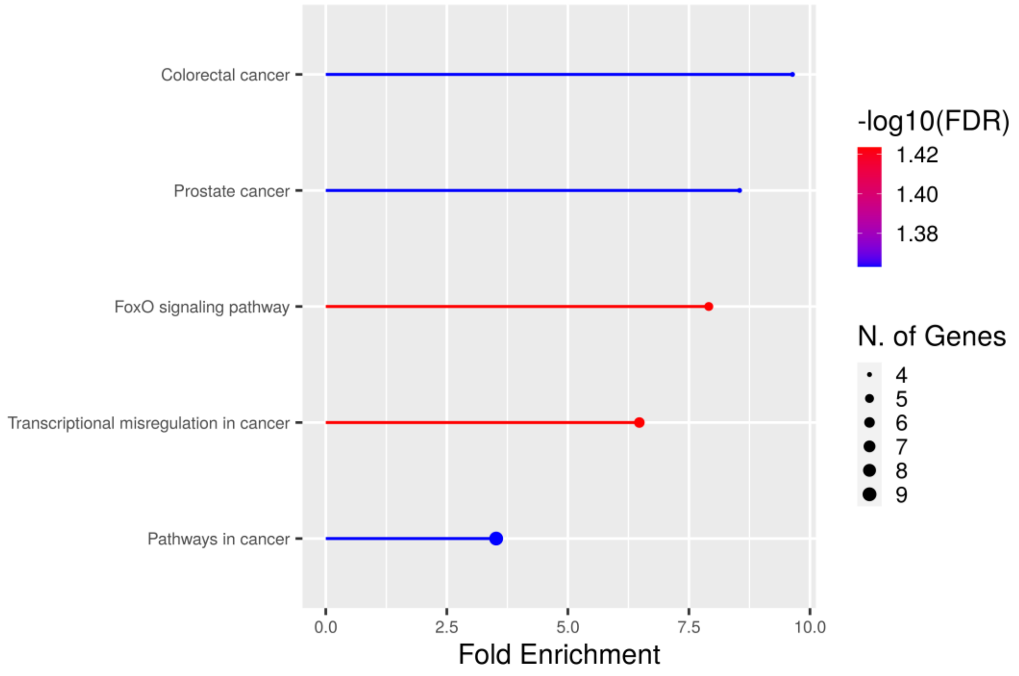
3.Results

Gene set enrichment for target genes of 2 miRNAs in the final model

MIMAT0005792 = hsa-miR-320b (3337 target) (left)
hsa-miR-320b: BCL9L, ZNF600, DCLRE1B, MAP7D1,..., CKS1B, EIF2A, SYNM, ACPP

MIMAT0022838 = hsa-miR-1185-1-3p (368 target) (right)
hsa-miR-1185-1-3p: TBK1, NR2F2, USP1, PMAIP,..., SESN3, SHOC2, SRI, ZFX, HAT1



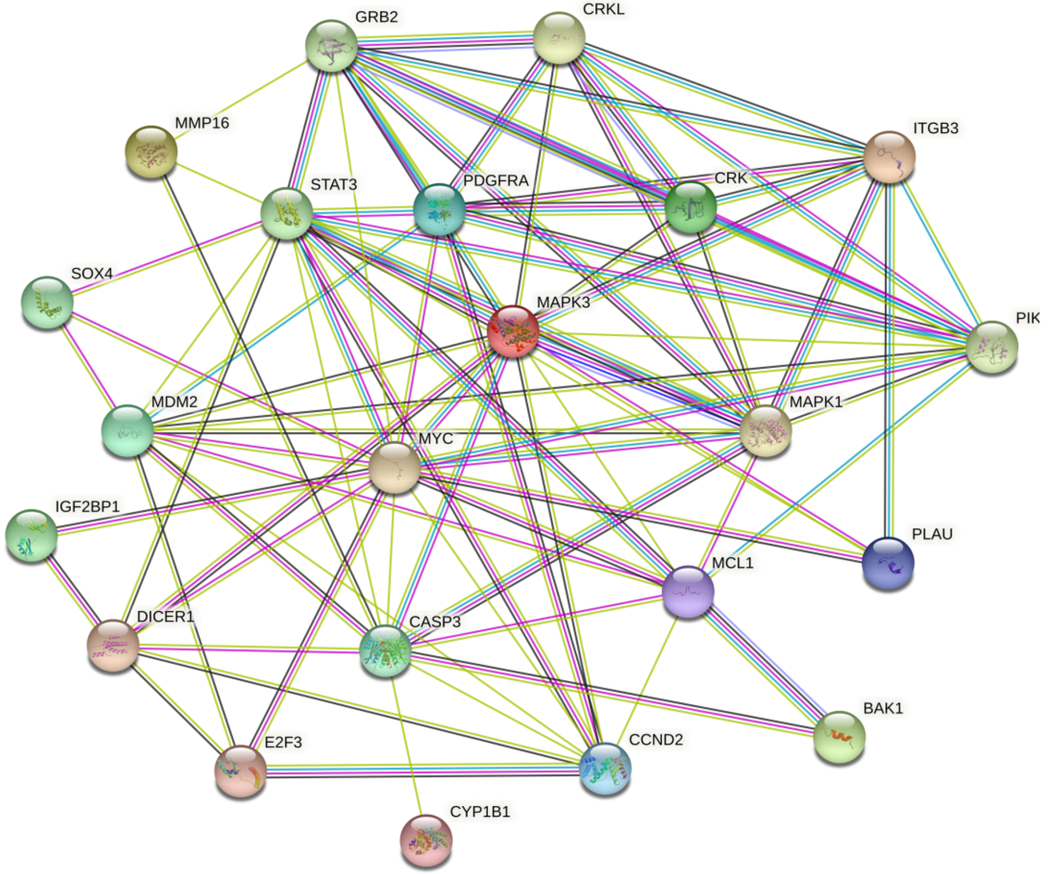


3.Results

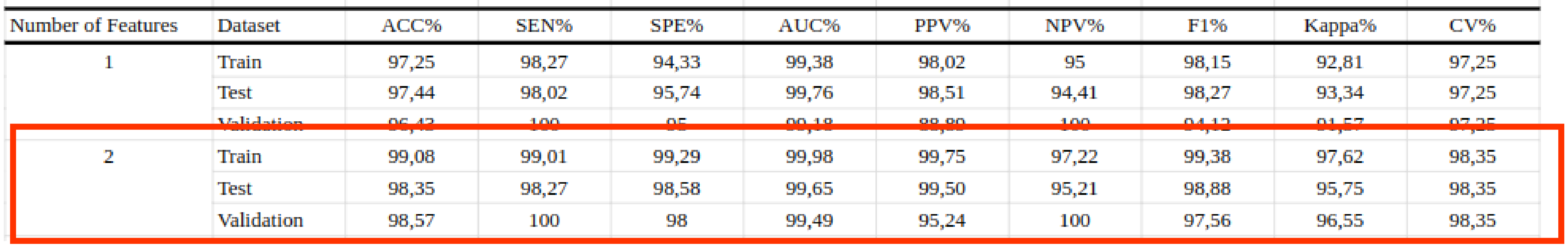
Protein -Protein Interaction Network for target genes of 2miRNAs in the final model

MIMAT0005792 = hsa-miR-320b (3337 target)

hsa-miR-320b: BCL9L, ZNF600, DCLRE1B, MAP7D1,..., CKS1B, EIF2A, SYNM, ACPP



4.Discussions



Number of Features	Dataset	ACC%	SEN%	SPE%	AUC%	PPV%	NPV%	F1%	Kappa%	CV%
1	Train	97,25	98,27	94,33	99,38	98,02	95	98,15	92,81	97,25
	Test	97,44	98,02	95,74	99,76	98,51	94,41	98,27	93,34	97,25
	Validation	96,43	100	95	99,18	88,89	100	94,12	91,57	97,25
2	Train	99,08	99,01	99,29	99,98	99,75	97,22	99,38	97,62	98,35
	Test	98,35	98,27	98,58	99,65	99,50	95,21	98,88	95,75	98,35
	Validation	98,57	100	98	99,49	95,24	100	97,56	96,55	98,35

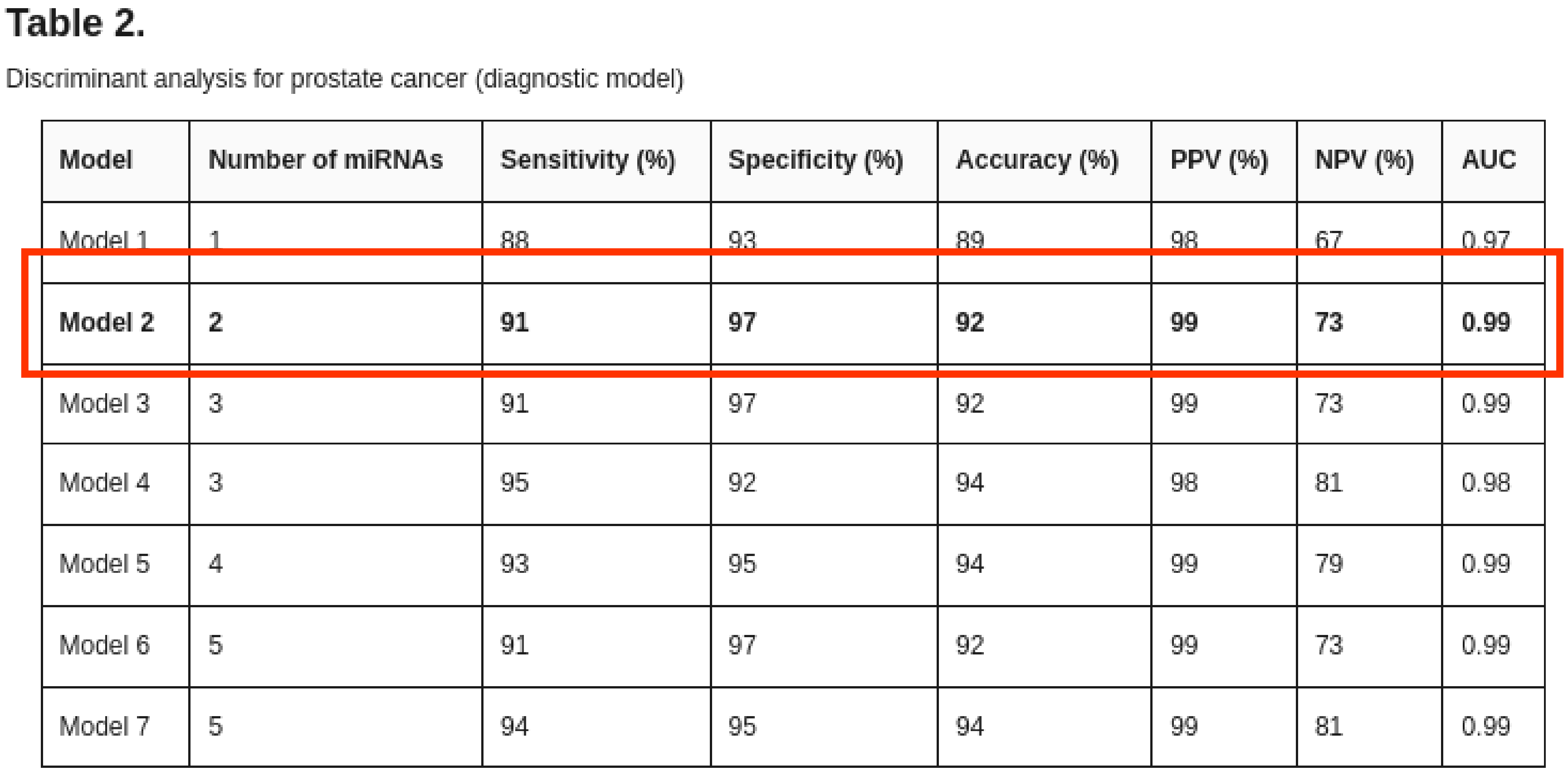
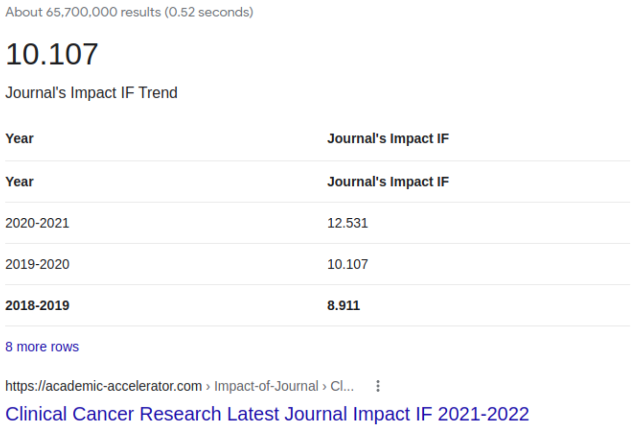


Table 2.

Discriminant analysis for prostate cancer (diagnostic model)

Model	Number of miRNAs	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)	AUC
Model 1	1	88	93	89	98	67	0.97
Model 2	**2**	**91**	**97**	**92**	**99**	**73**	**0.99**
Model 3	3	91	97	92	99	73	0.99
Model 4	3	95	92	94	98	81	0.98
Model 5	4	93	95	94	99	79	0.99
Model 6	5	91	97	92	99	73	0.99
Model 7	5	94	95	94	99	81	0.99



About 65,700,000 results (0.52 seconds)

10.107

Journal's Impact IF Trend

Year	Journal's Impact IF
Year	**Journal's Impact IF**
2020-2021	12.531
2019-2020	10.107
2018-2019	**8.911**

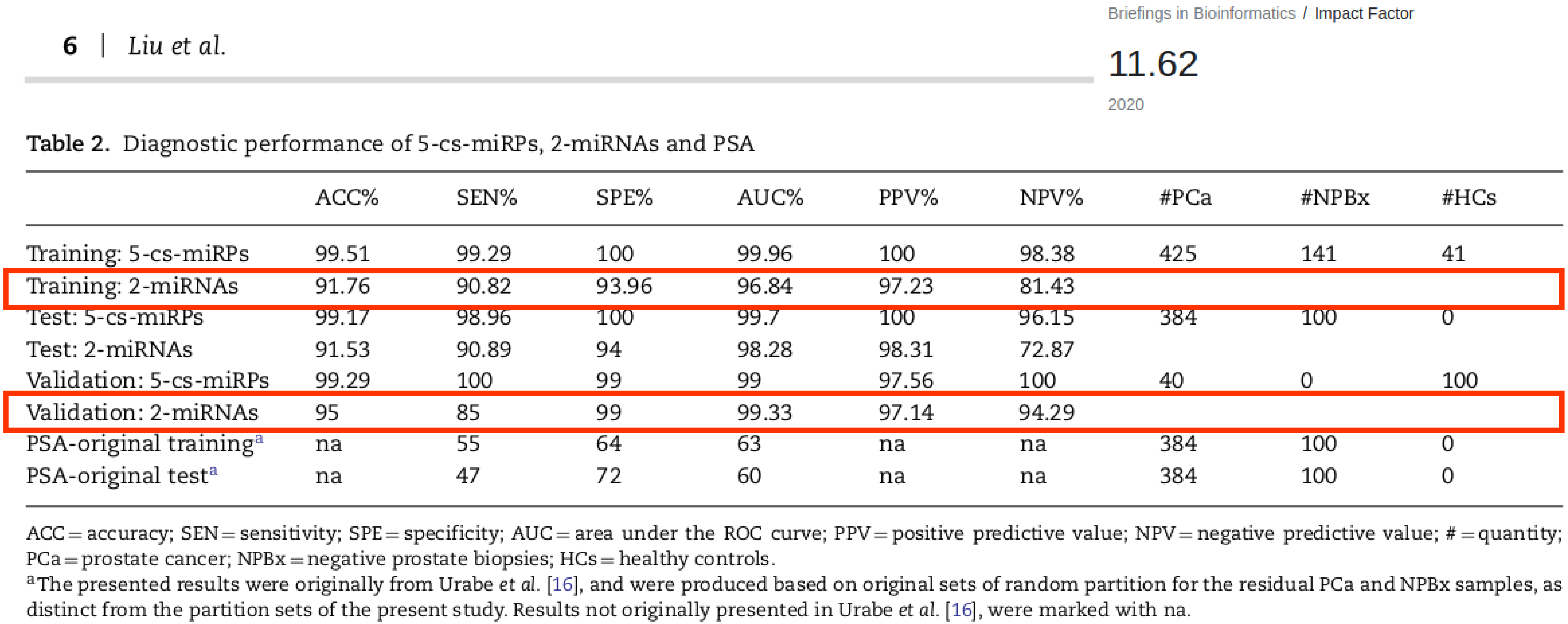
8 more rows

https://academic-accelerator.com › Impact-of-Journal › Cl...

Clinical Cancer Research Latest Journal Impact IF 2021-2022

https://doi.org/10.1158/1078-0432.CCR-18-2849

4.Discussions



6 | Liu et al.

Briefings in Bioinformatics / Impact Factor

11.62

2020

Table 2. Diagnostic performance of 5-cs-miRPs, 2-miRNAs and PSA

	ACC%	SEN%	SPE%	AUC%	PPV%	NPV%	#PCa	#NPBx	#HCs
Training: 5-cs-miRPs	99.51	99.29	100	99.96	100	98.38	425	141	41
Training: 2-miRNAs	91.76	90.82	93.96	96.84	97.23	81.43			
Test: 5-cs-miRPs	99.17	98.96	100	99.7	100	96.15	384	100	0
Test: 2-miRNAs	91.53	90.89	94	98.28	98.31	72.87			
Validation: 5-cs-miRPs	99.29	100	99	99	97.56	100	40	0	100
Validation: 2-miRNAs	95	85	99	99.33	97.14	94.29			
PSA-original training[a]	na	55	64	63	na	na	384	100	0
PSA-original test[a]	na	47	72	60	na	na	384	100	0

ACC = accuracy; SEN = sensitivity; SPE = specificity; AUC = area under the ROC curve; PPV = positive predictive value; NPV = negative predictive value; # = quantity; PCa = prostate cancer; NPBx = negative prostate biopsies; HCs = healthy controls.

[a] The presented results were originally from Urabe *et al.* [16], and were produced based on original sets of random partition for the residual PCa and NPBx samples, as distinct from the partition sets of the present study. Results not originally presented in Urabe *et al.* [16], were marked with na.

Our models is outperformed when we compared with original study (adjusted logistic regression) and the secondary study(variant SVM).

https://doi.org/10.1093/bib/bbaa111

PER SAPIENTIAM AD ASTRA

TTU

TAN TAO UNIVERSITY

AFTER THE SHOW, SHARE - LIKE - RETWEET!

THANKS

— FOR —

LISTENING