Supporting Information for "Predicting lncRNA–disease associations based on a dual-path feature extraction network with multiple sources of information integration"

Dengju Yao[1*], Binbin Zhang[1], Xiaojuan Zhan[1,2], Bo Zhang[1], Xiang Kui Li[1]

[1]School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China;

[2]College of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin, 150050, China;

*** Correspondence: Dengju Yao**

ydkvictory@hrbust.edu.cn;

1. Parameter analysis

1.1 GCN Layer Options

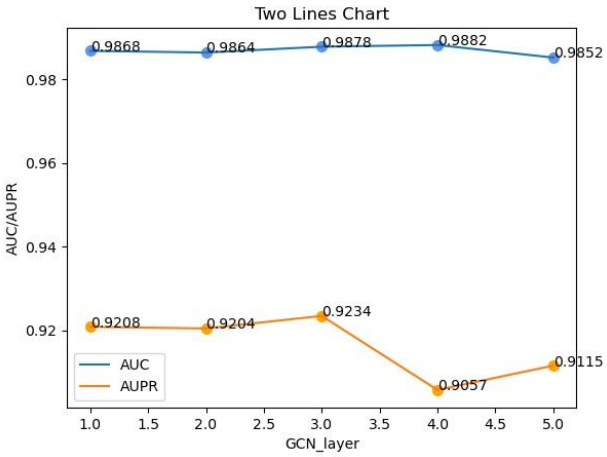

Figure S 1 GCN Layer Options by 5CV1

1.2 Optimization of Transformer parameters

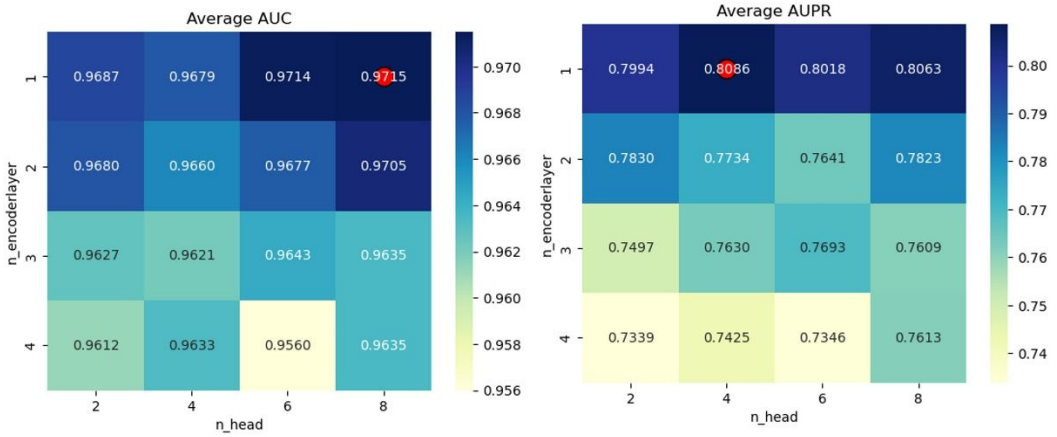

Figure S 2 Optimization of Transformer parameters by 5CV1

1.3 Output feature dimension selection

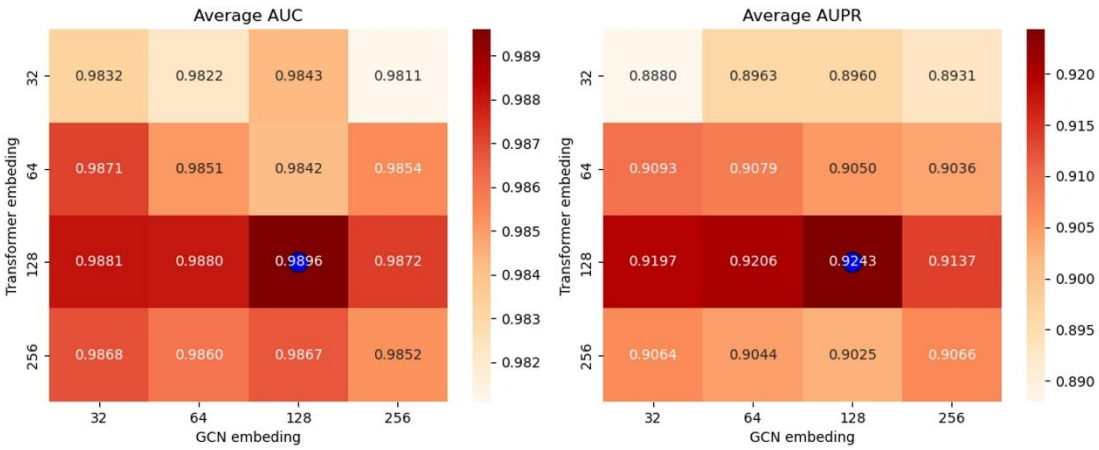

Figure S 3 Feature Size Selection for Dual Paths by 5CV1. The selection of feature sizes in the two paths, with the horizontal axis being the size of the topological features and the vertical axis being the size of the compensating features.

2. Classifier Comparison

Table S 1 Comparison of node classifier selection by 5CV1

	ACC	Specificity	Precision	Recall	MCC	F1-score
LR	0.9850	0.9977	0.7249	0.3137	0.4709	0.4379
GNB	0.8605	0.8598	0.1078	**0.8950**	0.2823	0.1925
SVM	0.9837	0.9983	0.7033	0.2125	0.3810	0.3265
RF	0.9876	0.9982	0.8234	0.4238	0.5856	0.5596
XGB	**0.9950**	**0.9993**	**0.9555**	0.7676	**0.8541**	**0.8512**

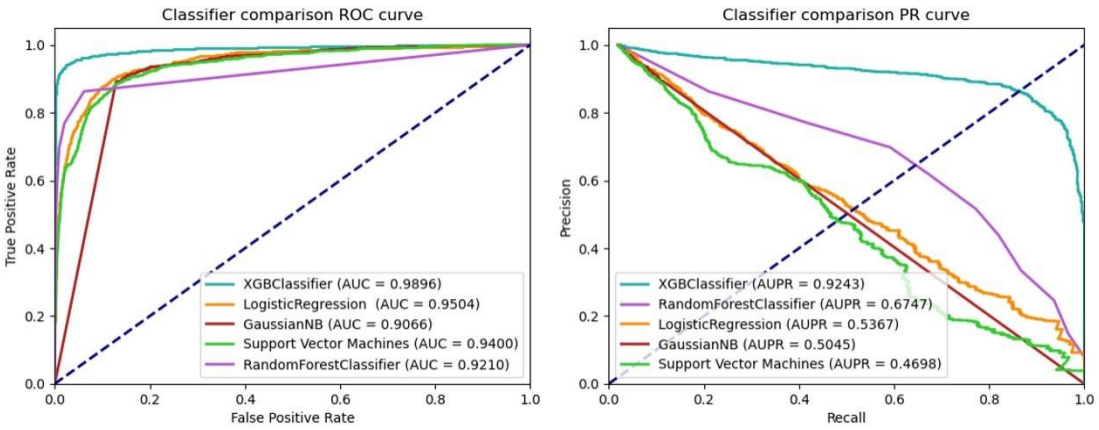

Figure S 4 AUC and AUPR for classifiers by 5CV1

3. Model Exploration

3.1 Node Aggregation Layer Technology Options

Table S 2 Node Aggregation Layer Technology Options by CV1 in dataset2

	PCA	ICA	RP	NAL
ACC	0.9916	0.9912	0.9917	**0.9950**
Specificity	0.9986	0.9989	0.9981	**0.9993**
Sensitivity	0.6223	0.5851	0.6248	**0.7676**
Precision	0.8966	0.9104	0.9020	**0.9555**
MCC	0.7432	0.7260	0.7470	**0.8541**
F1-score	0.7346	0.7124	0.7383	**0.8512**
AUC	0.9692	0.9679	0.9737	**0.9896**
AUPR	0.8104	0.8022	0.8248	**0.9243**

Comparison of different dimensionality reduction techniques on the structure of node feature extraction network with dual paths is carried out for classification using XGB classifier. The results show the superiority of NAL node aggregation layer.

3.2 Feature Extraction Network Substitution

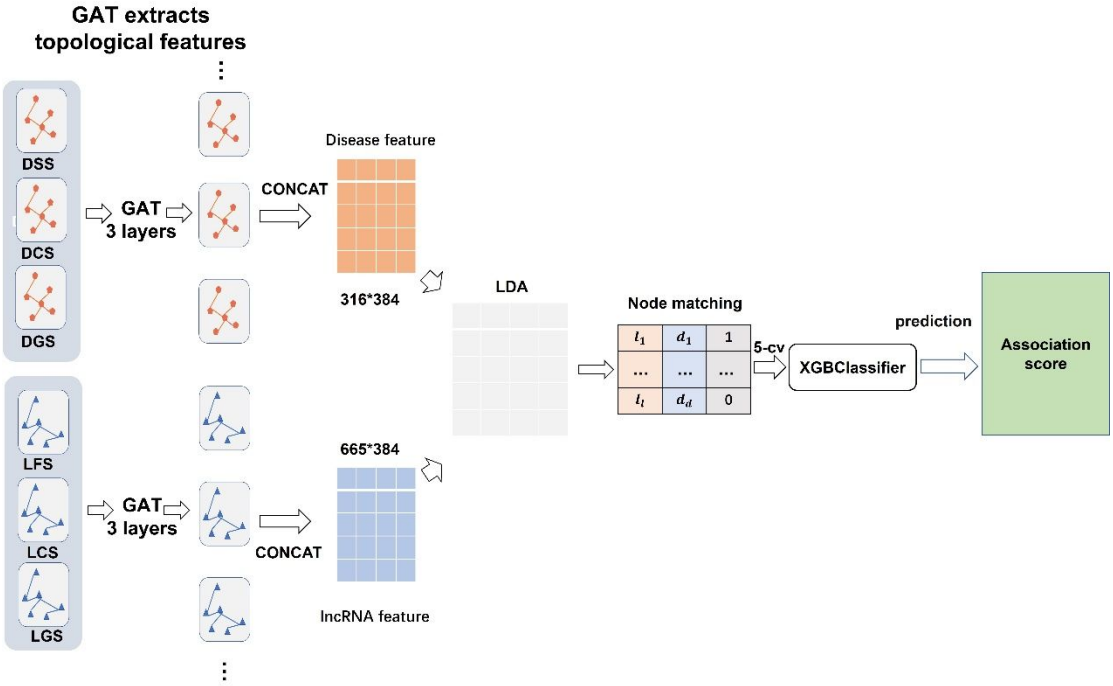

Figure S 5 Flowchart of topological feature extraction network for GAT

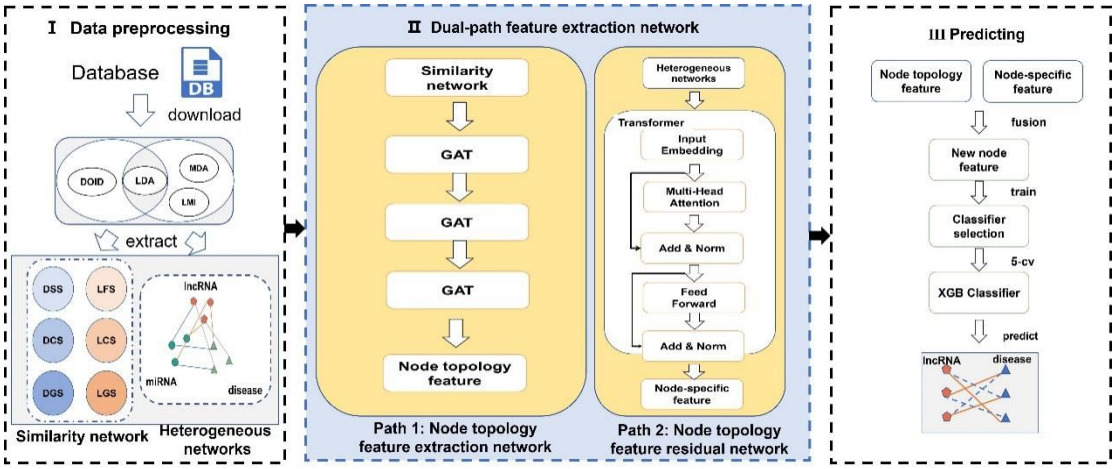

Figure S 6 Topological Feature Replacement for GAT Topological Feature Extraction Network Flowchart

In order to ensure the fairness of the experimental results, the GAT-based feature extraction model and the topological feature extraction network in the dual pathway use the same network topology for the disease view (DSS, DCS, DGS) and the lncRNA view (LFS, LCS, LGS) to explore the node features. The input and output node features are all 128-dimensional, and after feature splicing, the individual nodes become 384-dimensional. All of them are ultimately classified using the XGBoost classifier and result in the final prediction score. In order to further verify whether the node topology residual extraction network can alleviate the problem of disappearing node specificity brought by GAT, we replaced the topology feature extraction network in the dual-path topology extraction network with the graph attention topology feature extraction network. The results are shown in Table S 3.

Table S 3 DPFELDA vs. GAT by 5CV2

metric	Dual path feature extraction	GAT	GAT+ topological feature residual extraction network
ACC	**0.9942**	0.9847	0.9909
Specificity	**0.9991**	0.9972	0.9989
Sensitivity	**0.7349**	0.3239	0.5672
Precision	**0.9427**	0.6931	0.9134
Recall	**0.7349**	0.3239	0.5672
MCC	**0.8297**	0.4675	0.7159
F1-score	**0.8258**	0.4415	0.6698
AUC	**0.9888**	0.9319	0.9640
AUPR	**0.9222**	0.4951	0.7928

4.Wilcoxon test analysis

We did a Wilcoxon test analysis using the AUC and AUPR results of each comparison model in the test set on five occasions, which proved that due to the p-value being less than 0.05, the predictive performance of DPFELDA was significantly better than the methods compared.

Table S 4 Results of the paired Wilcoxon test comparing DPFELDA with all other methods on dataset1 and dataset2 by 5CV2

	p-value of AUC (D1)	p-value of AUPR (D1)	p-value of AUC(D2)	p-value of AUPR(D2)
HOPEXGB	2.52E-12	4.41E-10	1.78E-14	1.41E-14
SDLDA	1.23E-05	1.19E-06	1.85E-08	1.92E-12
GAEMCLDA	1.51E-11	1.11E-15	3.12E-12	6.43E-18
IPCARF	1.06E-06	1.46E-08	3.05E-08	1.7E-11
LDAformer	0.001089	8.27E-07	2.8E-09	6.69E-12
GCLMTP	1.07E-15	5.11E-14	1.35E-11	6.86E-16

Where D1 denotes dataset 1 and D2 denotes dataset 2.

5. Case study

5.1 Case study on dataset1

1 Breast cancer

Table S 5 breast cancer

Rank	lncRNA	Evidence	PMID	Rank	lncRNA	Evidence	PMID
1	MEG3	LncRNADisease	14602737	16	CCAT2	lnc2Cancer	28480695
2	HOTAIR	LncRNADisease	19182780	17	UCA1	LncRNADisease	16914571
3	MALAT1	LncRNADisease	18006640	18	LINC00583	Unknown	
4	PVT1	LncRNADisease	17908964	19	SOX2-OT	LncRNADisease	26409453
5	H19	LncRNADisease	28102845	20	ZFAS1	LncRNADisease	21460236
6	NEAT1	LncRNADisease	25417700	21	BCAR4	LncRNADisease	21506106
7	XIST	LncRNADisease	26637364	22	CCAT1	LncRNADisease	26464701
8	CDKN2B-AS1	LncRNADisease	34374638	23	HOTTIP	LncRNADisease	29415429
9	GAS5	LncRNADisease	22487937	24	LINC00472	LncRNADisease	29453409
10	BCYRN1	LncRNADisease	27277684	25	PANDAR	LncRNADisease	26927017
11	LSINCT5	lnc2Cancer	29785740	26	CASC16	Literature	36540803
12	AFAP1-AS1	lnc2Cancer	29974352	27	LINC-PINT	unknown	
13	KCNQ1OT1	LncRNADisease	21304052	28	MIR124-2HG	unknown	
14	LINC-ROR	LncRNADisease	26883251	29	CASC2	lnc2Cancer	31352515
15	LINC00675	unknown		30	MIR17HG	EVlncRNAs3	36943627

2 Gastric cancer

Table S 6 gastric cancer

Rank	lncRNA	Evidence	PMID	Rank	lncRNA	Evidence	PMID
1	MALAT1	LncRNADisease	24857172	16	AFAP1-AS1	LncRNADisease	28975981
2	MEG3	LncRNADisease	24006224	17	CASC2	LncRNADisease	27648142
3	CDKN2B-AS1	LncRNADisease	25636450	18	SPRY4-IT1	LncRNADisease	25835973
4	H19	LncRNADisease	27143813	19	NPTN-IT1	LncRNADisease	25674261
5	HOTAIR	LncRNADisease	27900563	20	KCNQ1OT1	LncRNADisease	25765901
6	GAS5	LncRNADisease	25959498	21	BANCR	LncRNADisease	26054683
7	UCA1	LncRNADisease	26718650	22	TUG1	LncRNADisease	27983921
8	NEAT1	LncRNADisease	28401449	23	WT1-AS	LncRNADisease	26449525
9	PVT1	LncRNADisease	25956062	24	TP53COR1	EVlncRNAs3	35947460
10	LSINCT5	lnc2Cancer	30127643	25	CCDC26	EVlncRNAs3	31166382
11	CCAT1	lnc2Cancer	31478245	26	HOTTIP	lnc2Cancer	27546609
12	MIR17HG	unknown		27	TUSC7	lnc2Cancer	25765901
13	XIST	LncRNADisease	27620004	28	NPSR1-AS1	unknown	
14	BCYRN1	lnc2Cancer	31652309	29	ZFAS1	lnc2Cancer	30999814
15	CCAT2	LncRNADisease v	27904778	30	GHET1	lnc2Cancer	28578256

3 Colorectal cancer

Table S 7 colorectal cancer

Rank	lncRNA	Evidence	PMID	Rank	lncRNA	Evidence	PMID
1	MALAT1	lncRNADisease	21503572	16	ZFAS1	lncRNADisease	26506418
2	MEG3	lncRNADisease	25636452	17	BANCR	lncRNADisease	25013510
3	NPSR1-AS1	lncRNADisease	34406628	18	HOTTIP	lncRNADisease	26678886
4	LINC00583	unknown		19	CRNDE	lncRNADisease	22393467
5	HOTAIR	lncRNADisease	21862635	20	UCA1	lncRNADisease	26238511
6	CAHM	lncRNADisease	24799664	21	GAS5	lncRNADisease	27391432
7	TUSC7	lncRNADisease	23680400	22	H19	lncRNADisease	22427002
8	NEAT1	lncRNADisease	26314847	23	TP53COR1	lncRNADisease	24573322
9	TUG1	lncRNADisease	28302487	24	CCAT1	lncRNADisease	23594791
10	NCRUPAR	lncRNADisease	25119598	25	SNHG4	lncRNADisease	33744866
11	PVT1	lncRNADisease	26990997	26	KCNQ1OT1	lnc2Cancer	30997746
12	XIST	lncRNADisease	28730777	27	HOTAIRM1	LncRNADisease	27307307
13	CASC16	unknown		28	LSINCT5	lnc2Cancer	25526476
14	CCAT2	lnc2Cancer v3.0	31558855	29	PRNCR1	lnc2Cancer	26530130
15	GHET1	lncRNADisease	27131316	30	HULC	lnc2Cancer	27496341

5.2 Case study on dataset2

1 Breast cancer

Table S 8 breast cancer

Rank	lncRNA	Evidence	PMID	Rank	lncRNA	Evidence	PMID
1	LINC00668	lnc2Cancer	32117742	16	LSINCT3	LncRNADisease	20214974
2	EPB41L4A-AS1	lnc2Cancer	30959550	17	LINC01234	LncRNADisease	27338266
3	DLEU1	LncRNADisease	26416600	18	DRAIC	LncRNADisease	25288503
4	LINC00673	lnc2Cancer	30094100	19	LINC01016	LncRNADisease	26426411
5	KCNK15-AS1	LncRNADisease	25929808	20	NKILA	LncRNADisease	25759022
6	MEG3	LncRNADisease	22393162	21	STXBP5-AS1	LncRNADisease	27338266
7	LINC00473	lnc2Cancer	30848493	22	LINC00460	lnc2Cancer	31308741
8	SNHG15	LncRNADisease	29217194	23	FENDRR	lnc2Cancer	29559798
9	PCAN-1	LncRNADisease	27322459	24	CCAT1	LncRNADisease	26464701
10	SPRY4-IT1	LncRNADisease	25742952	25	PCAN-4	LncRNADisease	27322459
11	LINC00339	LncRNADisease	29453409	26	LINC01296	lnc2Cancer	29559798
12	LSINCT10	LncRNADisease	20214974	27	LINC00665	lnc2Cancer	32271427
13	BANCR	lnc2Cancer	29565494	28	DANCR	LncRNADisease	27716745
14	MT1JP	lnc2Cancer	32039825	29	RP11-434D9.1	LncRNADisease	26910840
15	ADARB2-AS1	LncRNADisease	26929647	30	SOX2	LncRNADisease	25006803

2 Gastric cancer

Table S 9 gastric cancer

Rank	lncRNA	Evidence	PMID	Rank	lncRNA	Evidence	PMID
1	PCGEM1	lnc2Cancer	31421977	16	GAS5	lnc2Cancer	31530437
2	HCP5	lnc2Cancer	32357145	17	LINC00941	lnc2Cancer	30723491
3	MHRT	lnc2Cancer	31273599	18	LINC00675	lnc2Cancer	29107103
4	LINC00665	lnc2Cancer	31736127	19	MIF-AS1	lnc2Cancer	30238562
5	HIF1A-AS2	lnc2Cancer	25686741	20	MIR100HG	lnc2Cancer	30886062
6	LBX2-AS1	lnc2Cancer	32351330	21	SNHG15	lnc2Cancer	26662309
7	FAM83H-AS1	lnc2Cancer	31493939	22	LINC00628	lnc2Cancer	27272474
8	FALEC	lnc2Cancer	30984243	23	PRNCR1	lnc2Cancer	26206497
9	FEZF1-AS1	lnc2Cancer	31785996	24	SNHG16	lnc2Cancer	30854107
10	LINC00629	lnc2Cancer	31674022	25	ZFAS1	lnc2Cancer	30999814
11	OIP5-AS1	lnc2Cancer	32147682	26	TP53TG1	lnc2Cancer	27821766
12	SPRY4-IT1	lnc2Cancer	31330497	27	ADAMTS9-AS2	lnc2Cancer	32160650
13	DANCR	lnc2Cancer	31002130	28	MVIH	lnc2Cancer	31773716
14	LINC01234	lnc2Cancer	32011780	29	TRPM2-AS	lnc2Cancer	32123162
15	MIAT	lnc2Cancer	32274858	30	LINC00668	lnc2Cancer	27036039

3 Colorectal cancer

Table S 10 colorectal cancer

Rank	lncRNA	Evidence	PMID	Rank	lncRNA	Evidence	PMID
1	LEF1-AS1	lnc2Cancer	32248974	16	FEZF1-AS1	lnc2Cancer	29914894
2	DANCR	lnc2Cancer	32159208	17	91H	lnc2Cancer	30978169
3	NKILA	lnc2Cancer	31423284	18	DLEU1	lnc2Cancer	30098595
4	LUADT1	lnc2Cancer	29762830	19	MEG3	lnc2Cancer	32219064
5	R05532	lnc2Cancer	25421768	20	DUXAP10	lnc2Cancer	28779166
6	RPL34-AS1	lnc2Cancer	24908062	21	PRNCR1	lnc2Cancer	24330491
7	ST3GAL6-AS1	lnc2Cancer	30613961	22	MIAT	lnc2Cancer	31567876
8	LINC00473	lnc2Cancer	30126852	23	LINC01503	lnc2Cancer	30542444
9	MIR100HG	lnc2Cancer	31814848	24	SNHG16	lnc2Cancer	30962265
10	SPRY4-IT1	lnc2Cancer	28099409	25	LINC00675	lnc2Cancer	29524886
11	FENDRR	lnc2Cancer	31724220	26	TI21327	None	None
12	LINC00472	lnc2Cancer	29488624	27	FAM83H-AS1	lnc2Cancer	29434883
13	lncAGER	lnc2Cancer	32031046	28	LOC100292680	None	None
14	CRNDE-h	lnc2Cancer	27042112	29	IGKV	None	None
15	ZFAS1	lnc2Cancer	30250022	30	TINCR	lnc2Cancer	30853664

4 hepatocellular carcinoma

Table S 11 hepatocellular carcinoma

Rank	lncRNA	Evidence	PMID	Rank	lncRNA	Evidence	PMID
1	MT1JP	lnc2Cancer	26909858	16	SOX2-OT	LncRNADisease	26097588
2	OIP5-AS1	lnc2Cancer	32042127	17	BANCR	LncRNADisease	26758762
3	TCL6	lnc2Cancer	32020591	18	WT1-AS	LncRNADisease	26462627
4	LINC01296	lnc2Cancer	30988624	19	RP51014016.1	LncRNADisease	25556502
5	CCEPR	LncRNADisease	27427851	20	P8725	LncRNADisease	25900874
6	LINC00628	lnc2Cancer	30740671	21	DUXAP10	lnc2Cancer	30996112
7	CPS1-IT1	LncRNADisease	27248828	22	GPC3-AS1	LncRNADisease	27573079
8	ENST00000501583	LncRNADisease	24876753	23	uc001ncr	LncRNADisease	26674525
9	MIR31HG	lnc2Cancer	30176933	24	LINC00339	lnc2Cancer	31239716
10	P16984	LncRNADisease	25900874	25	LINC00665	LncRNADisease	27499103
11	CDKN2B-AS1	LncRNADisease	25966845	26	SBF2-AS1	lnc2Cancer	30115383
12	DLX6-AS1	lnc2Cancer v3.0	29145165	27	FAM83H-AS1	lnc2Cancer	31599410
13	CCAT1	LncRNADisease	25884472	28	LINC01018	LncRNADisease	25512078
14	P23099	LncRNADisease	25900874	29	P9745	LncRNADisease	25900874
15	DUXAP8	lnc2Cancer v3.0	32022476	30	FEZF1-AS1	lnc2Cancer	29957463

5.3 Survivability analysis on dataset1

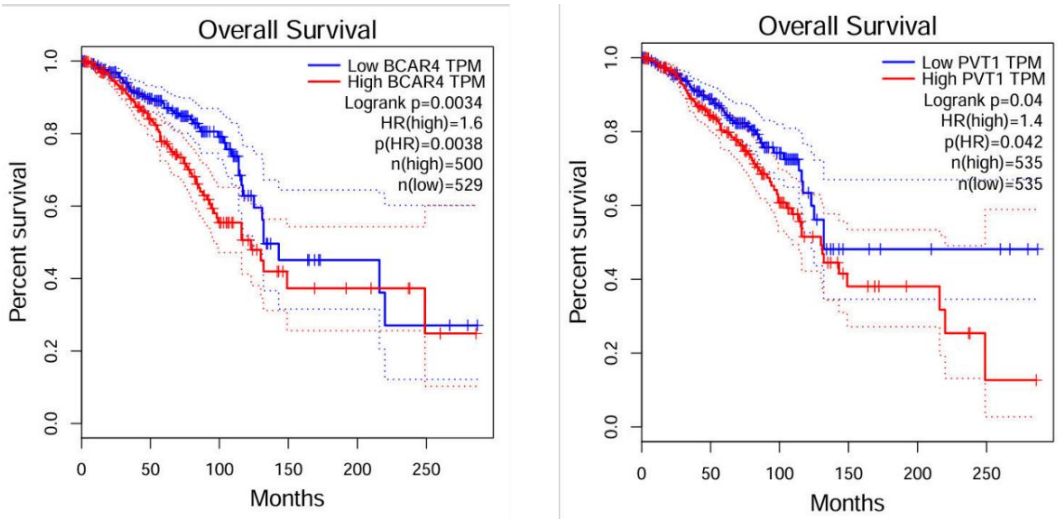

Figure S 7 Survivability analysis of breast cancer on dataset1

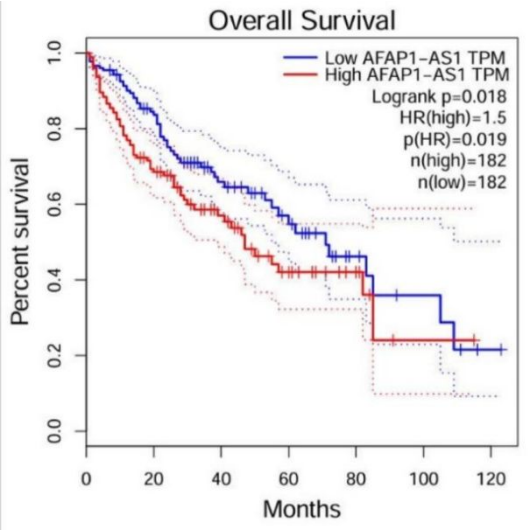

Figure S 8 Survivability analysis of hepatocellular carcinoma cancer on dataset1

5.4 Survivability analysis on dataset2

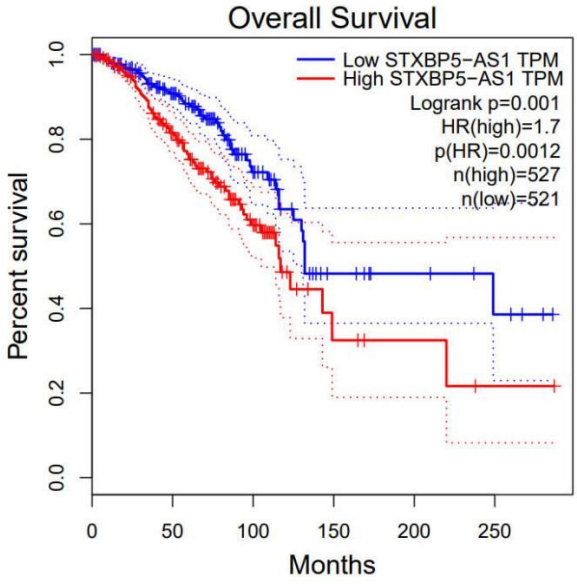

Figure S 9 Survivability analysis of breast cancer on dataset2

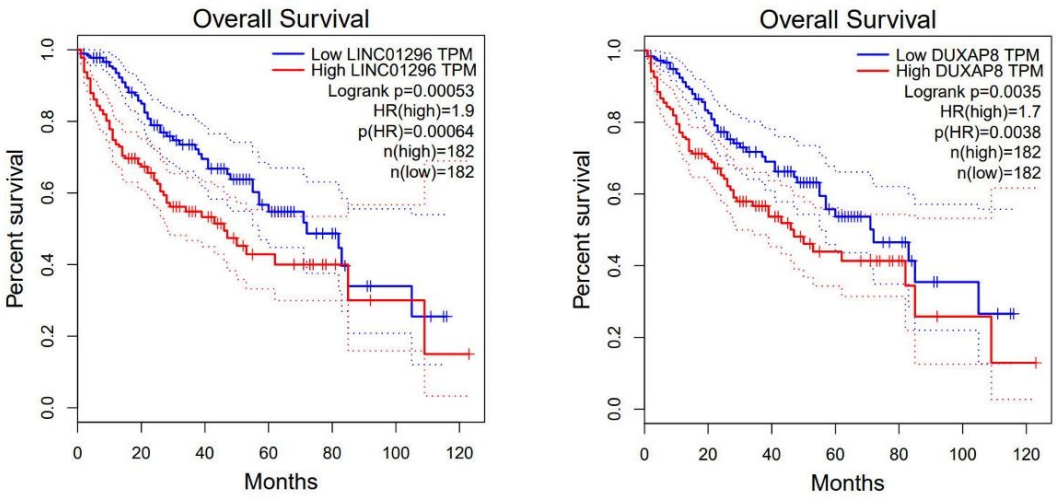

Figure S 10 Survivability analysis of hepatocellular carcinoma cancer on dataset2

6. Robustness experiments

Table S 12 Robustness experiments by 5-CV2

	Dataset1	Dataset2	Dataset3	Dataset4
ACC	0.9956	0.9942	0.9962	0.9888
precision	0.9796	0.9427	0.9773	0.9435
Recall	0.8565	0.7349	0.8846	0.7399
MCC	0.9139	0.8297	0.9279	0.8203
F1-score	0.9139	0.8259	0.9286	0.8293
AUC	0.9967	0.9888	0.9978	0.9860
AUPR	0.9671	0.9222	0.9763	0.9125