1. **What I have done:**

**(1).**

Did multiclass classification on the cancer types with features set as sbs signatures, using the softmax to construct the softmax logistic regression to perform the training and testing of the model, and extracted weight of each sbs signatures in each cancer types and gene types. **the result of the classification accuracy of all the fold’s validation dataset for cancer is of (5-fold cross validation):**

The classification is based on ['ACC', 'BLCA', 'BRCA', 'CESC', 'CHOL', 'COAD', 'DLBC', 'ESCA', 'GBM', 'HNSC', 'KICH', 'KIRC', 'KIRP', 'LAML', 'LGG', 'LIHC', 'LUAD', 'LUSC', 'MESO', 'OV', 'PAAD', 'PCPG', 'PRAD', 'READ', 'SKCM', 'STAD', 'TGCT', 'THCA', 'THYM', 'UCEC', 'UCS', 'UVM']

32 of cancers

and ['MEN1', 'ATRX', 'CTNNB1', 'TP53', 'PRKAR1A', 'RXRA', 'CUL1', 'NFE2L2', 'STAG2', 'EP300', 'FAT1', 'FOXA1', 'ATM',

'ELF3', 'FOXQ1', 'ARID1A', 'ERBB3', 'PIK3CA', 'ASXL2', 'KRAS', 'TXNIP', 'ZFP36L1', 'HRAS', 'RHOB', 'CREBBP',

'NRAS', 'ERBB2', 'KANSL1', 'KDM6A', 'PSIP1', 'FGFR3', 'CDKN2A', 'SF1', 'GNA13', 'ERCC2', 'DIAPH2', 'SF3B1',

'PTEN', 'RB1', 'KMT2D', 'CDKN1A', 'RHOA', 'TSC1', 'SPTAN1', 'KLF5', 'RBM10', 'FBXW7', 'KMT2C', 'TBX3', 'CDKN1B',

'NF1', 'PTPRD', 'CBFB', 'CHD4', 'PIK3R1', 'GATA3', 'CTCF', 'MAP3K1', 'AKT1', 'MAP2K4', 'CDH1', 'GPS2', 'CASP8',

'NCOR1', 'BRCA1', 'RUNX1', 'HLA-B', 'SMAD4', 'NOTCH1', 'POLRMT', 'STK11', 'TGFBR2', 'MAPK1', 'LATS1', 'IDH1', 'EPHA2',

'PBRM1', 'BAP1', 'TGIF1', 'APC', 'ACVR2A', 'AMER1', 'PCBP1', 'BRAF', 'SOX9', 'TCF7L2', 'ZFP36L2', 'SMAD2', 'GNAS', 'CARD11',

'CD70', 'TNFAIP3', 'TMSB4X', 'BTG2', 'HIST1H1E', 'PIM1', 'CD79B', 'MYD88', 'B2M', 'NSD1', 'PTCH1', 'ZNF750', 'TCF12', 'EGFR',

'LZTR1', 'SPTA1', 'KEL', 'PDGFRA', 'GABRA6', 'CYLD', 'MYH9', 'FLNA', 'HUWE1', 'AJUBA', 'KEAP1', 'RAC1', 'ARID2', 'CUL3', 'RASA1',

'HLA-A', 'SETD2', 'MTOR', 'VHL', 'KDM5C', 'KIF1A', 'TCEB1', 'SMARCB1', 'MET', 'NF2', 'SMC1A', 'KIT', 'U2AF1', 'PTPDC1', 'NPM1',

'DNMT3A', 'PTPN11', 'ASXL1', 'IDH2', 'FLT3', 'WT1', 'CIC', 'ZBTB20', 'FUBP1', 'NIPBL', 'MAX', 'SMARCA4', 'ZCCHC12', 'PLCG1',

'NUP133', 'ALB', 'BRD7', 'APOB', 'TSC2', 'AXIN1', 'XPO1', 'WHSC1', 'CREB3L3', 'IL6ST', 'DHX9', 'RPS6KA3', 'EEF1A1', 'MGA',

'RIT1', 'FGFR2', 'ARHGAP35', 'LATS2', 'CDK12', 'ZNF133', 'RNF43', 'EEF2', 'RET', 'CSDE1', 'EPAS1', 'MED12', 'ZMYM3', 'SPOP',

'DACH1', 'COL5A1', 'PPP6C', 'MAP2K1', 'CDK4', 'RQCD1', 'MECOM', 'DDX3X', 'GNA11', 'PTMA', 'PPM1D', 'EIF1AX', 'NUP93', 'GTF2I',

'CHD3', 'MSH6', 'BCOR', 'ARID5B', 'CCND1', 'ACVR1', 'PDS5B', 'DICER1', 'SIN3A', 'MYCN', 'ZFHX3', 'FOXA2', 'ATR', 'KMT2B',

'PIK3R2', 'TAF1', 'RPL22', 'PPP2R1A', 'ZMYM2', 'SCAF4', 'INPPL1', 'CTNND1', 'SOS1', 'RRAS2', 'SOX17', 'ATF7IP', 'RFC1',

'BCL2L11', 'MAP3K4', 'ZBTB7B', 'PLCB4', 'GNAQ', 'SRSF2', 'CYSLTR2']

total 224 genes

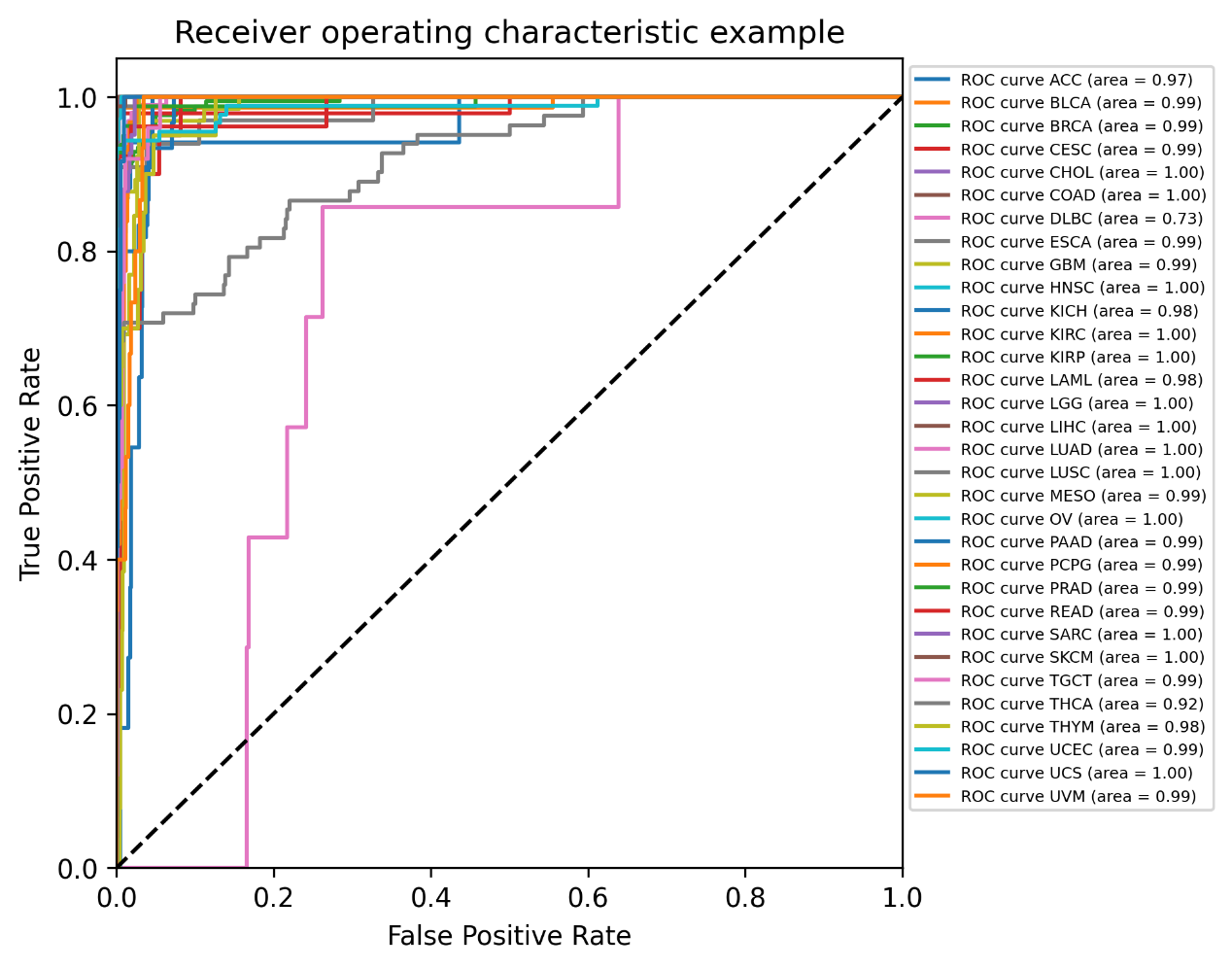
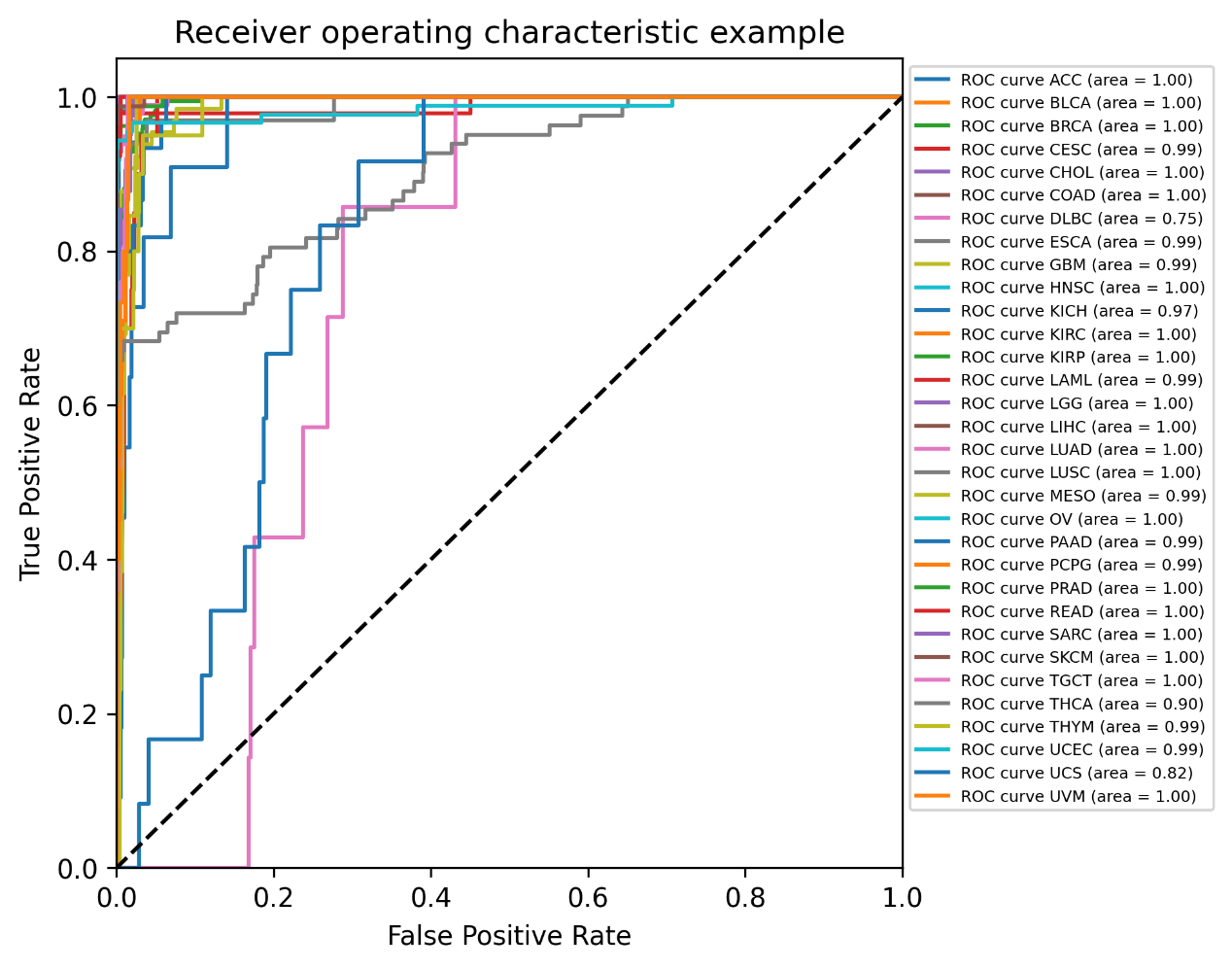
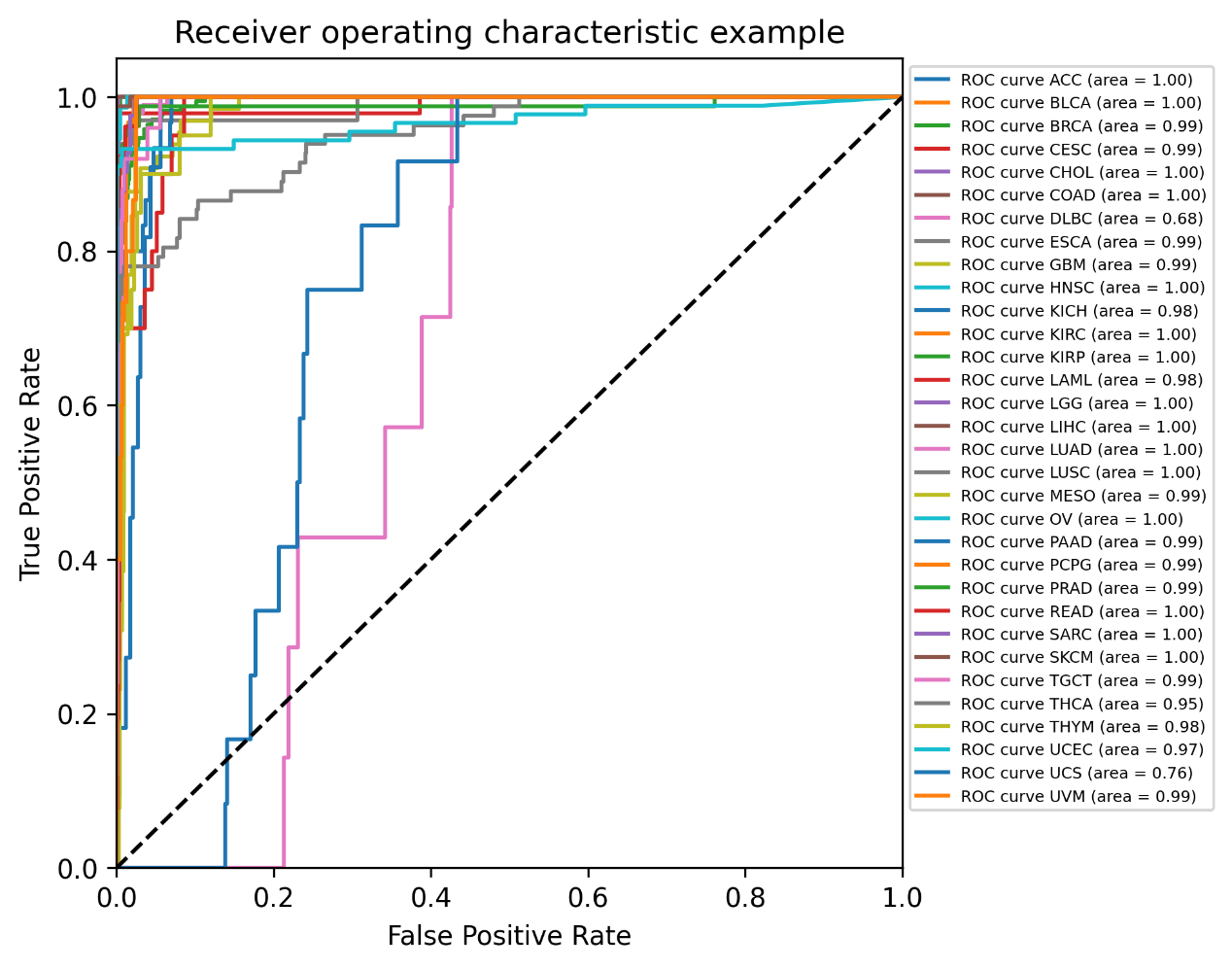
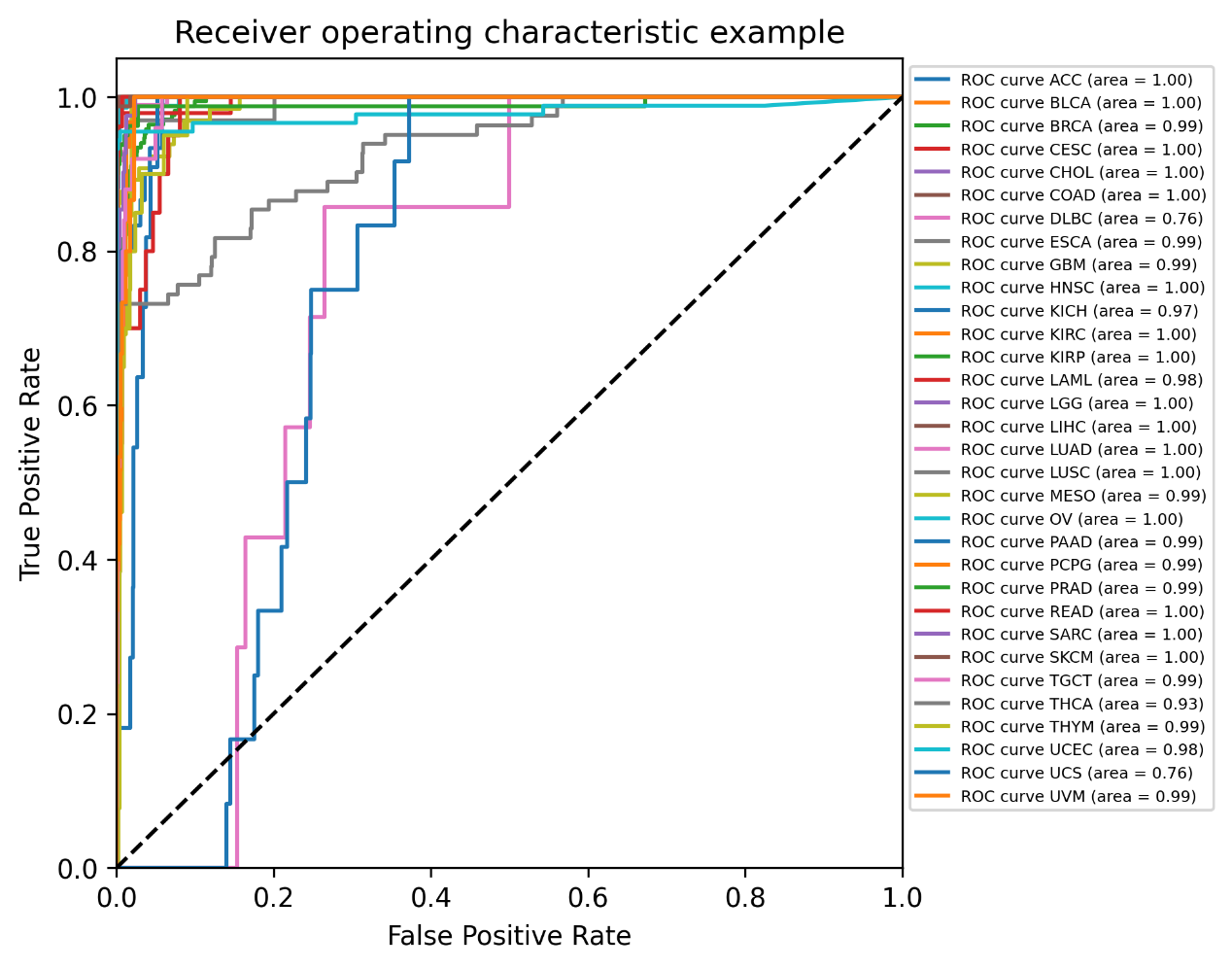
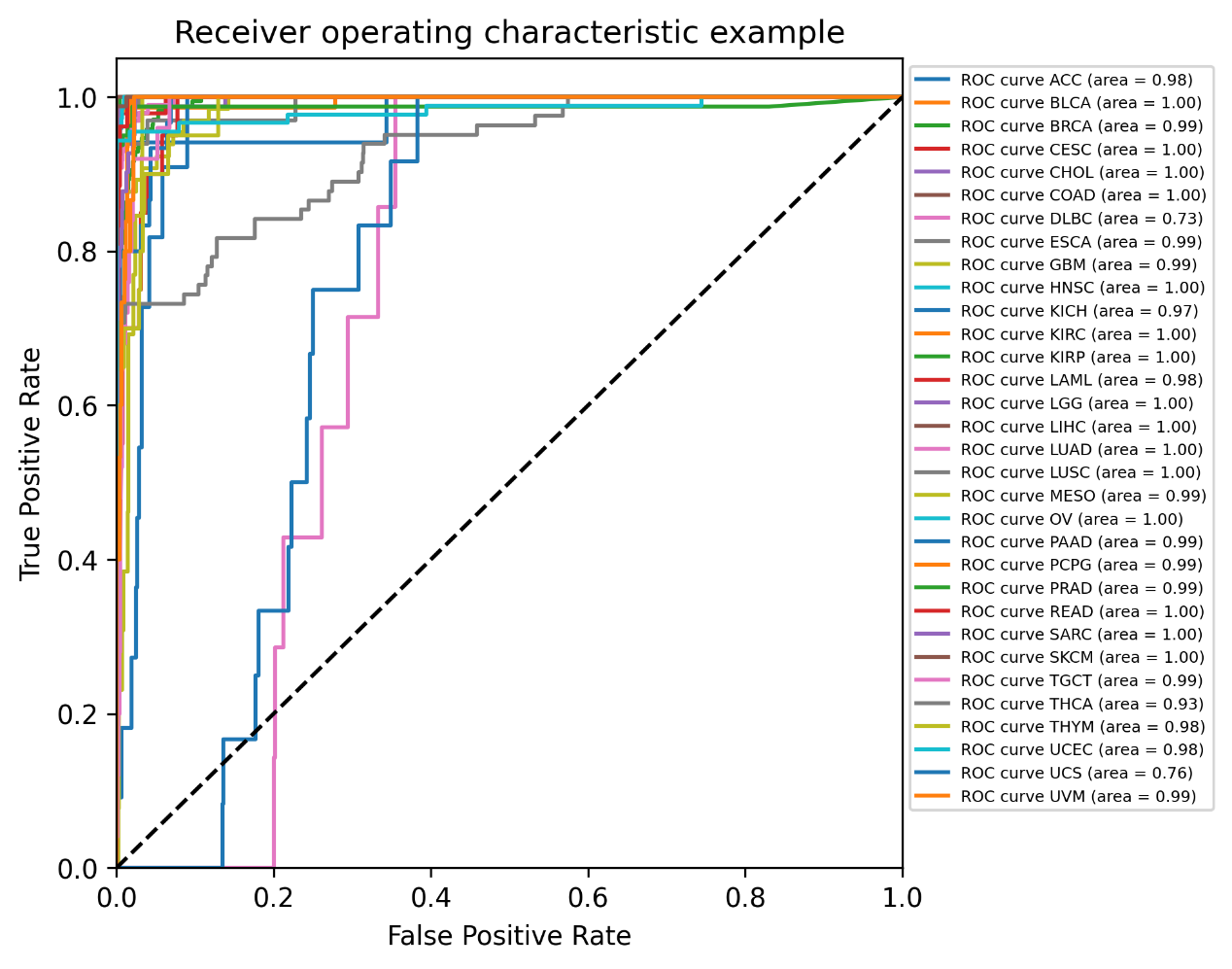
The 5-fold cross validation has 5 testing result, they are:

[0.8283208020050126,0.8890977443609023,**0.8953634085213033**,0.7901002506265664, 0.8865914786967418]

The validation accuracies for 5-fold cross validation are:

[0.8522316043425814,**0.8685162846803377**,0.8552472858866104,0.8480096501809409, 0.8413751507840772]

The classification report for the classification status of 5 validation sets are shown below:



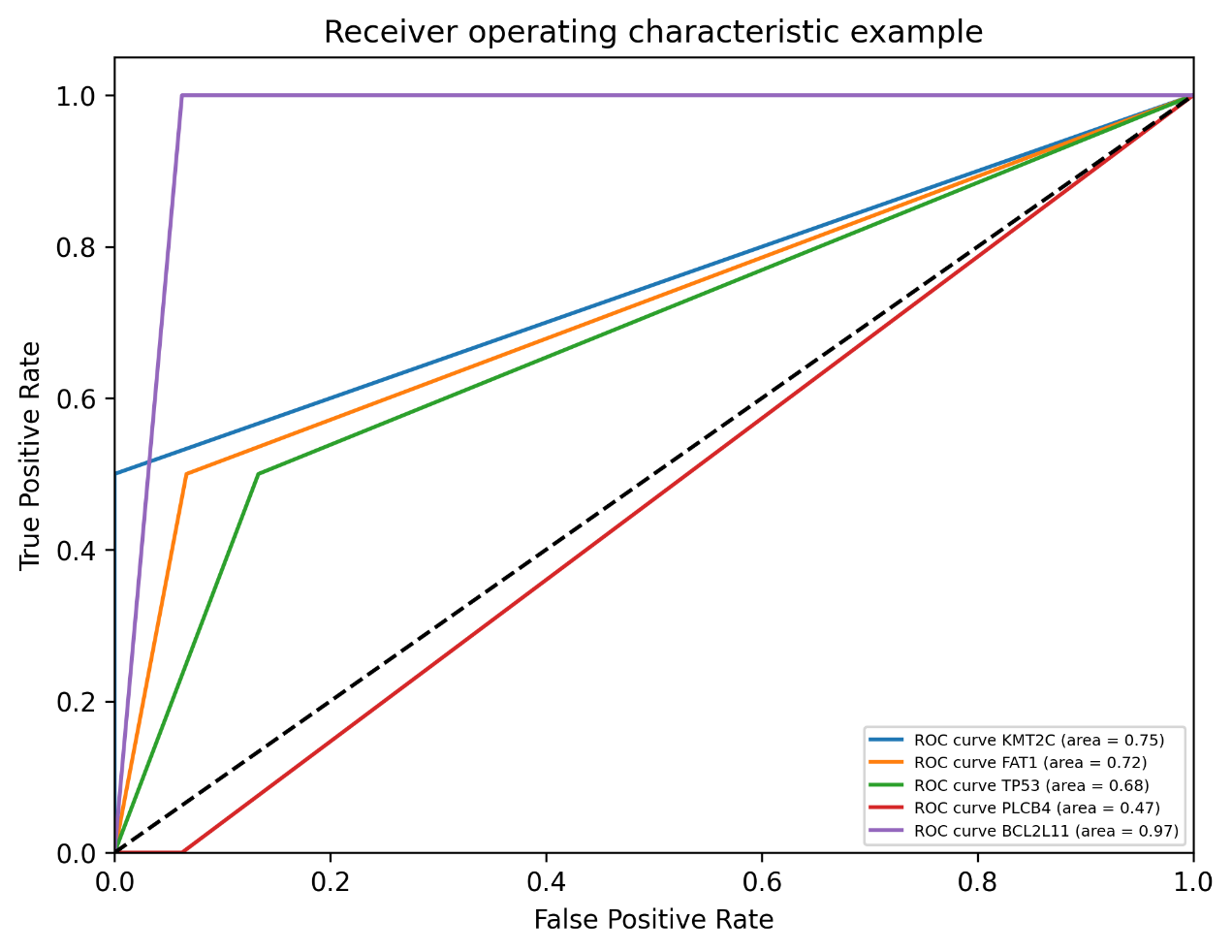
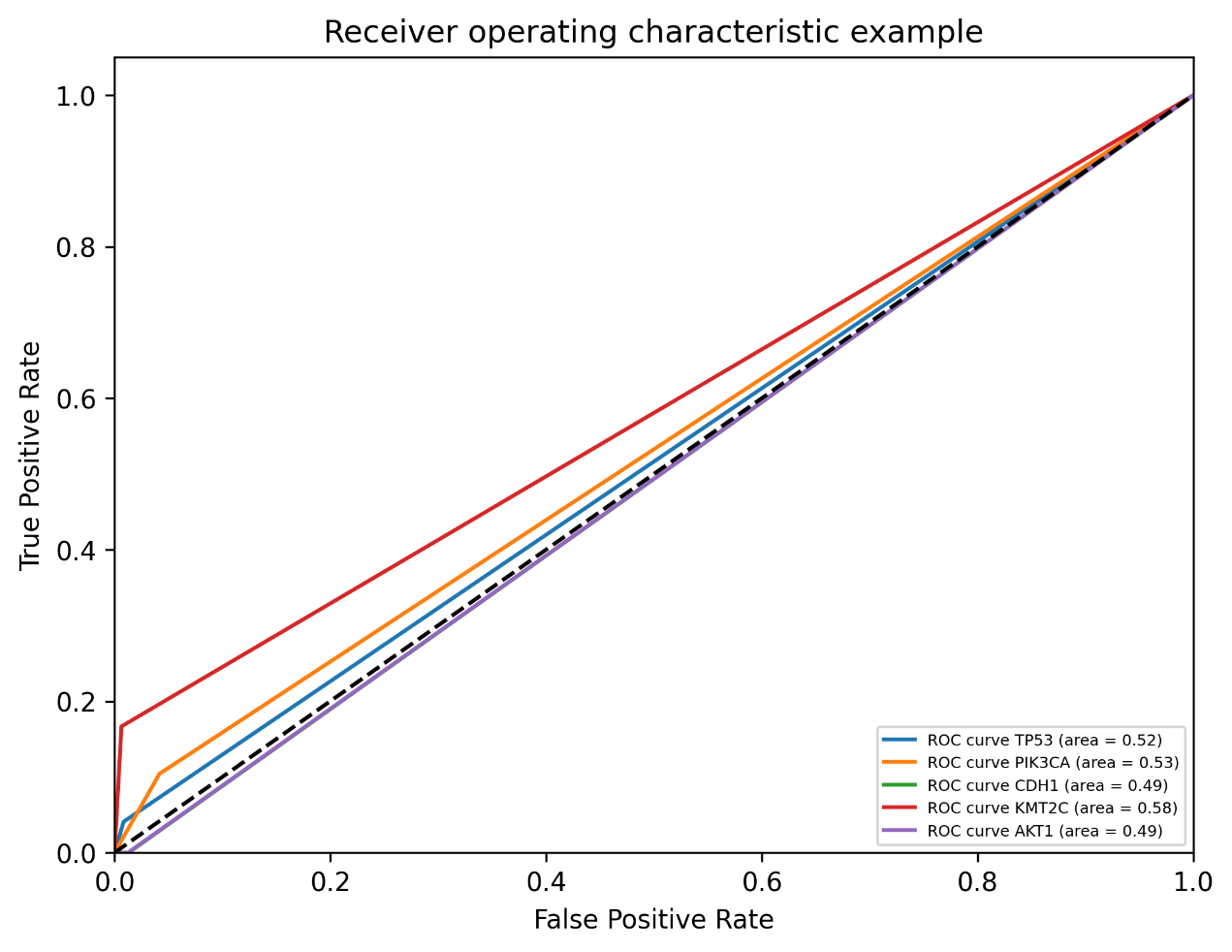
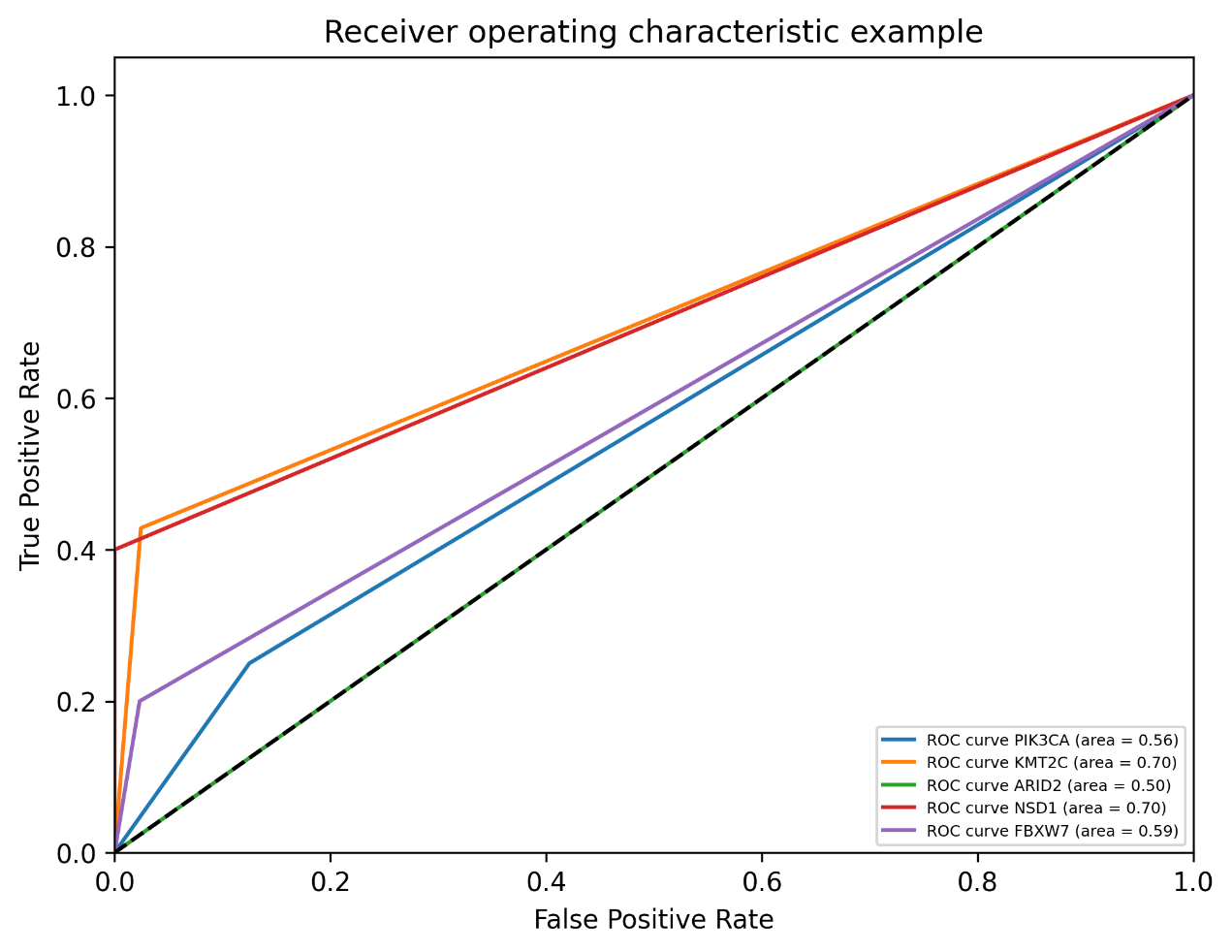
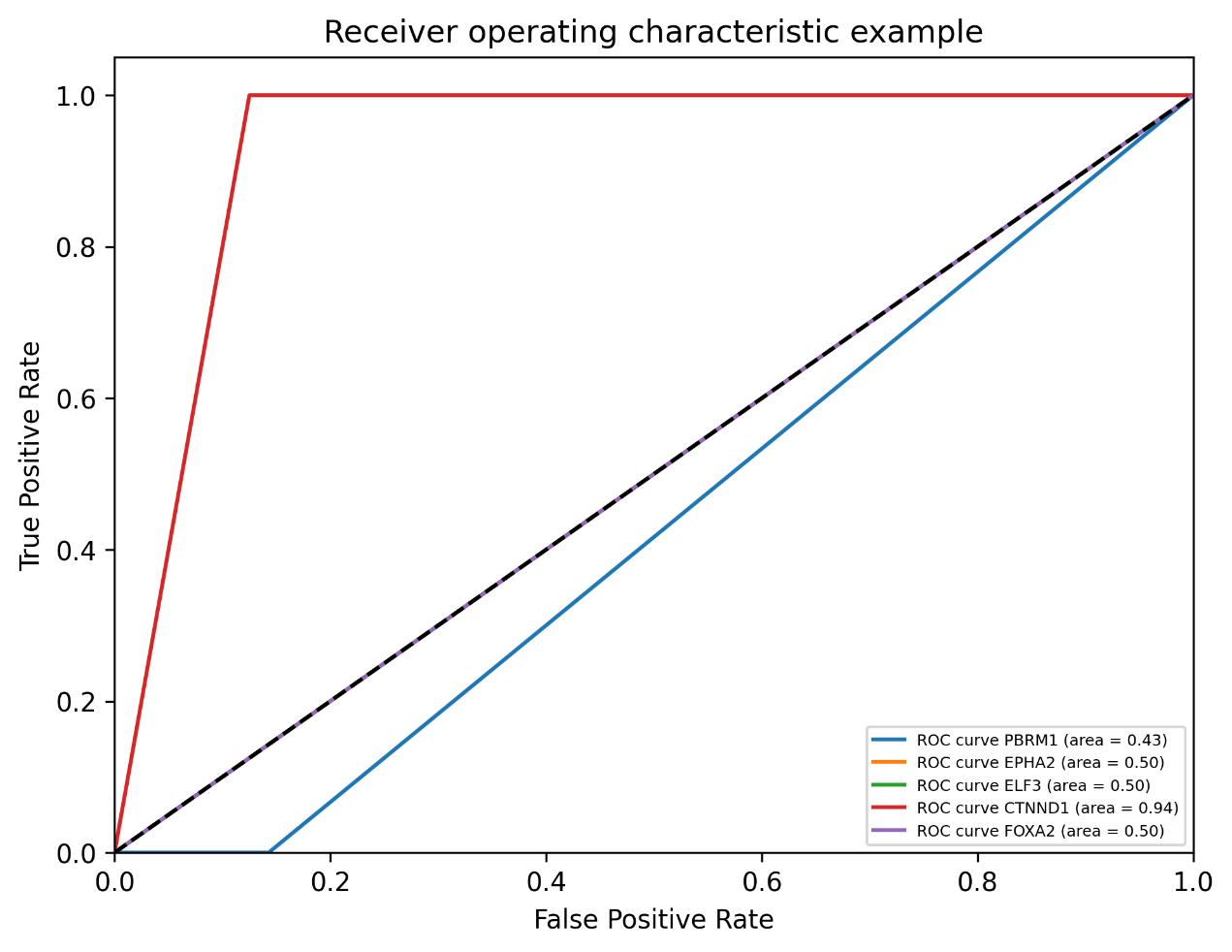
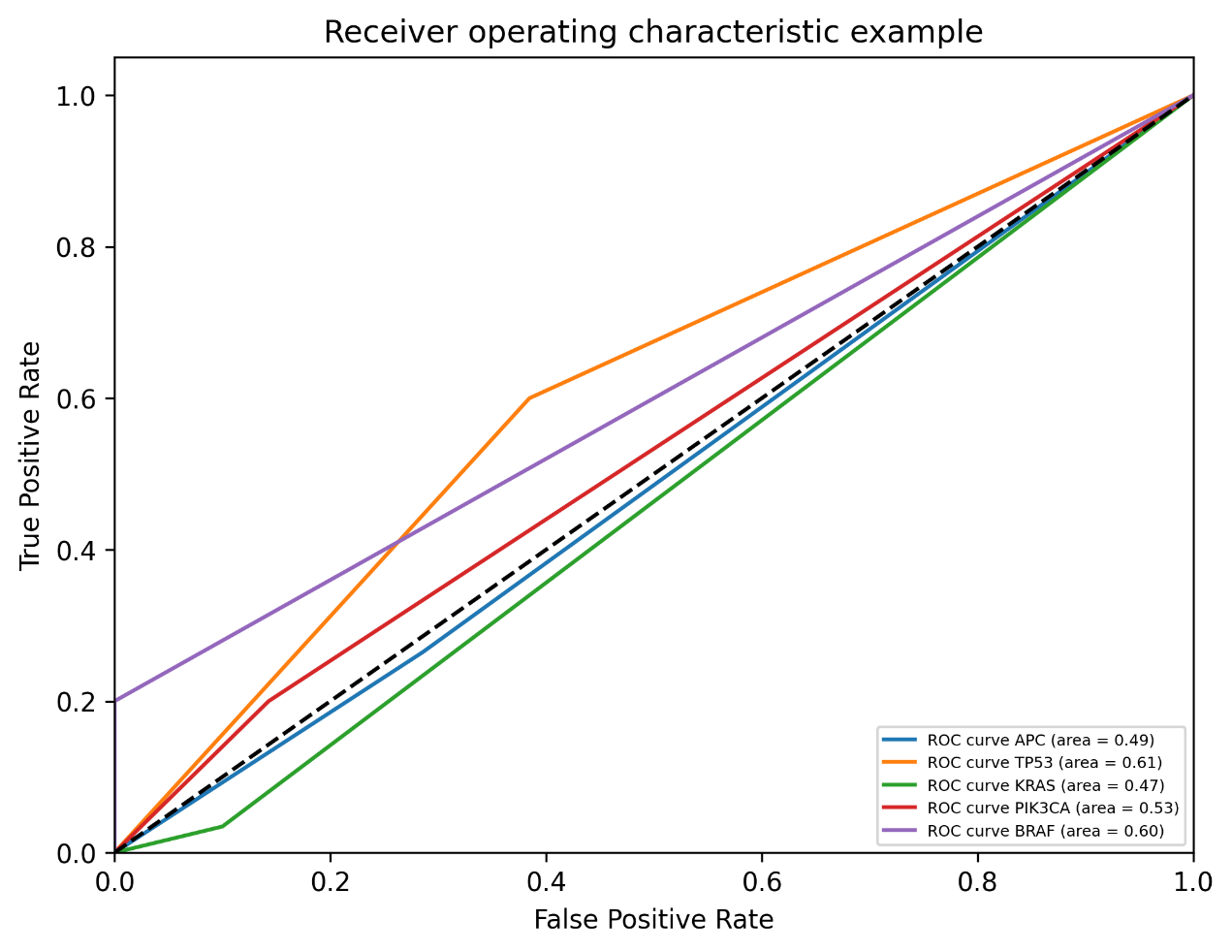
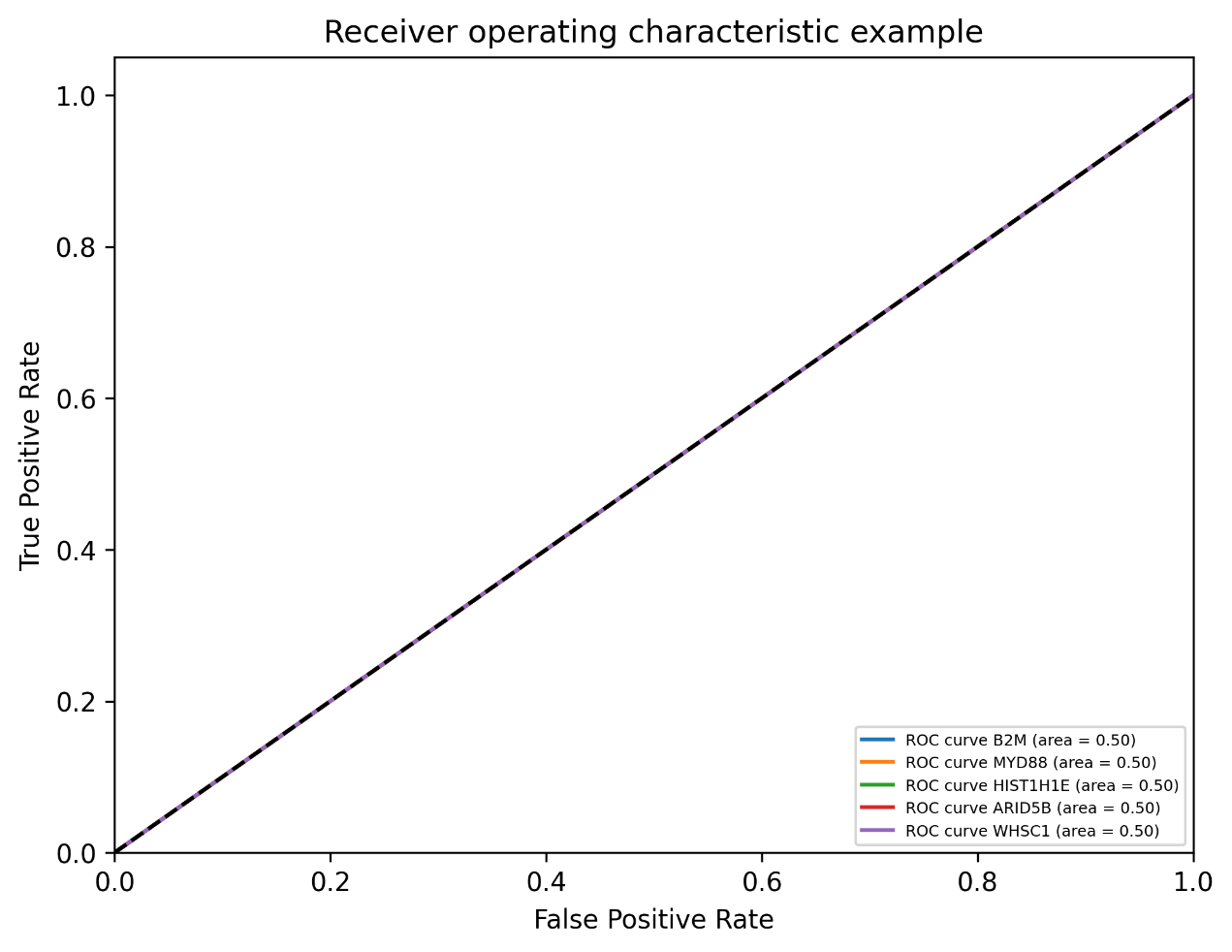
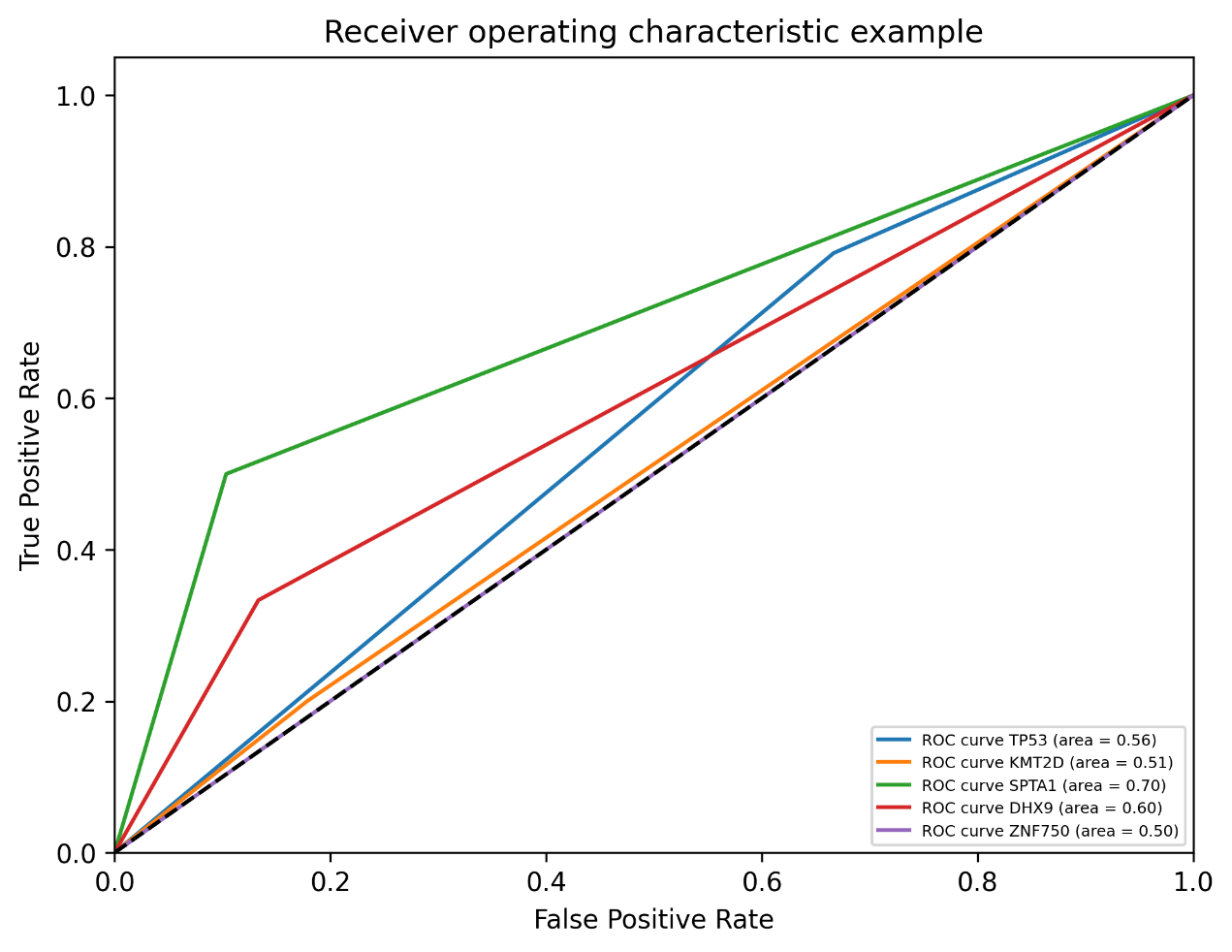
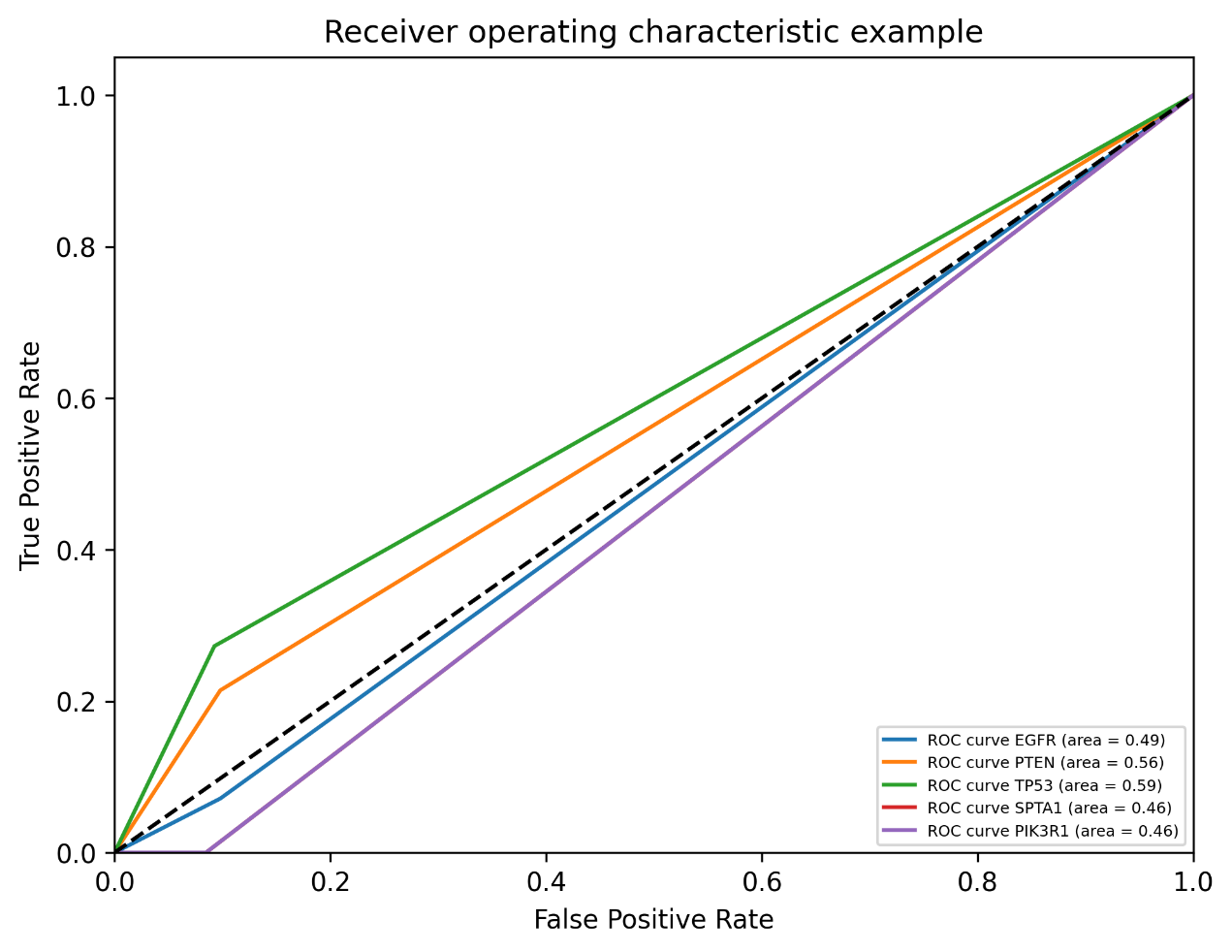
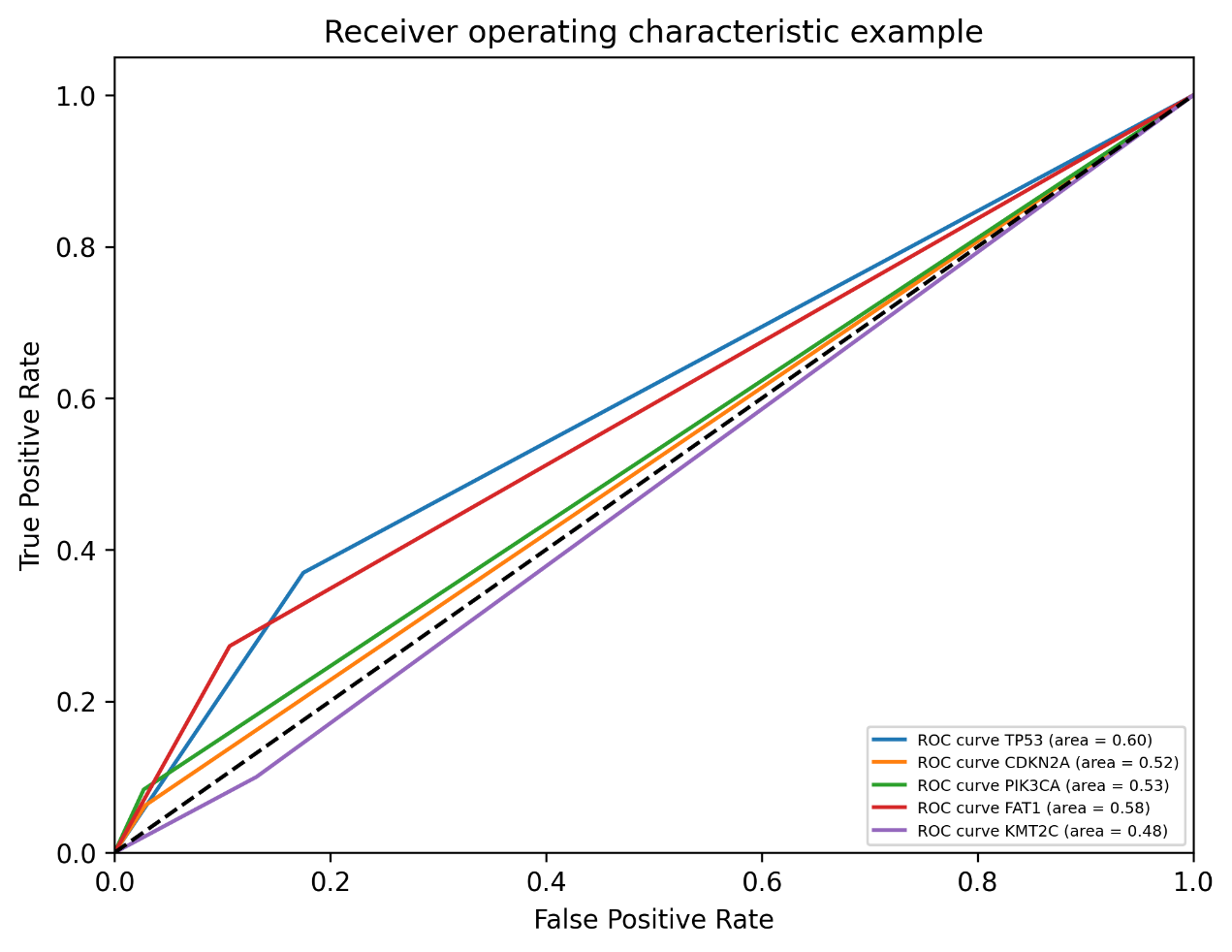
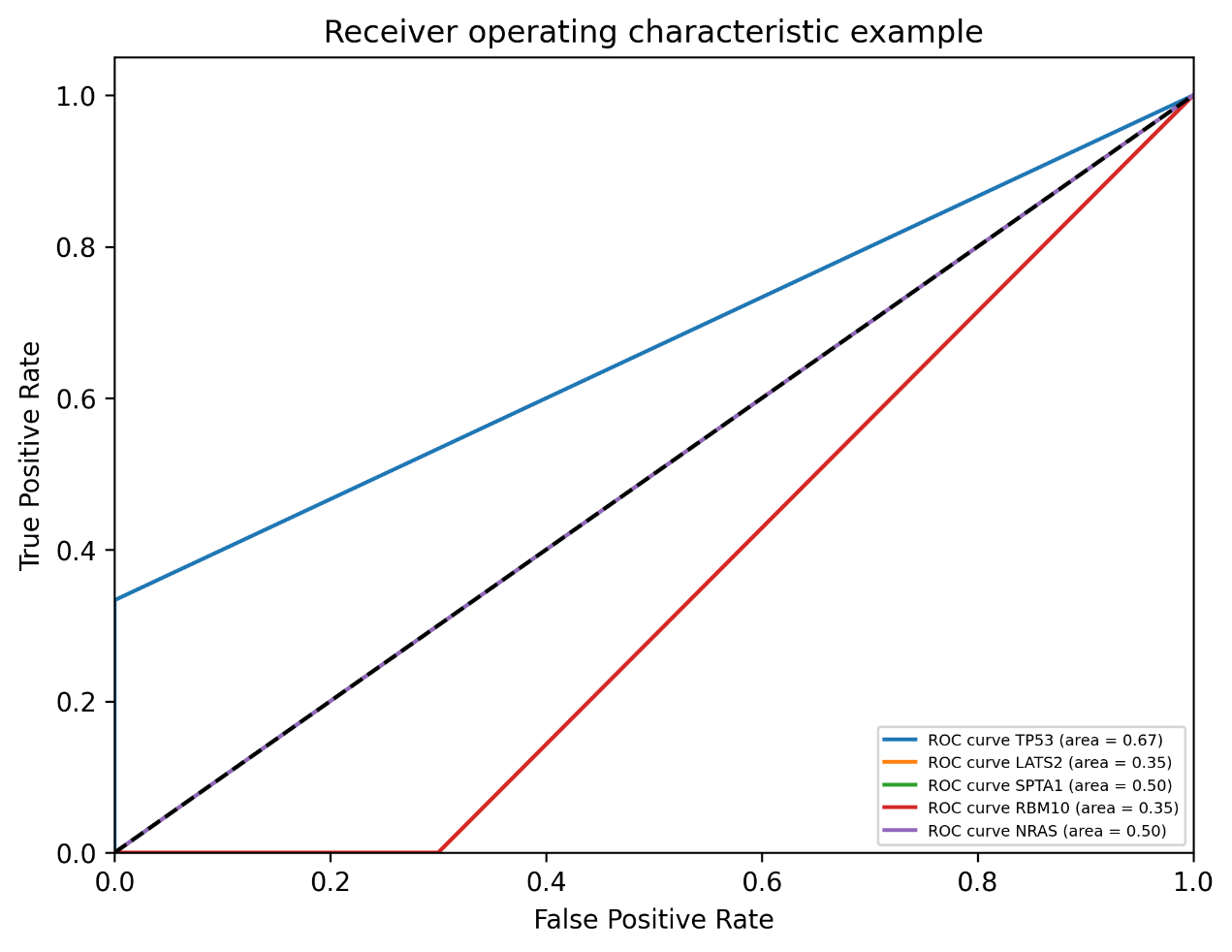
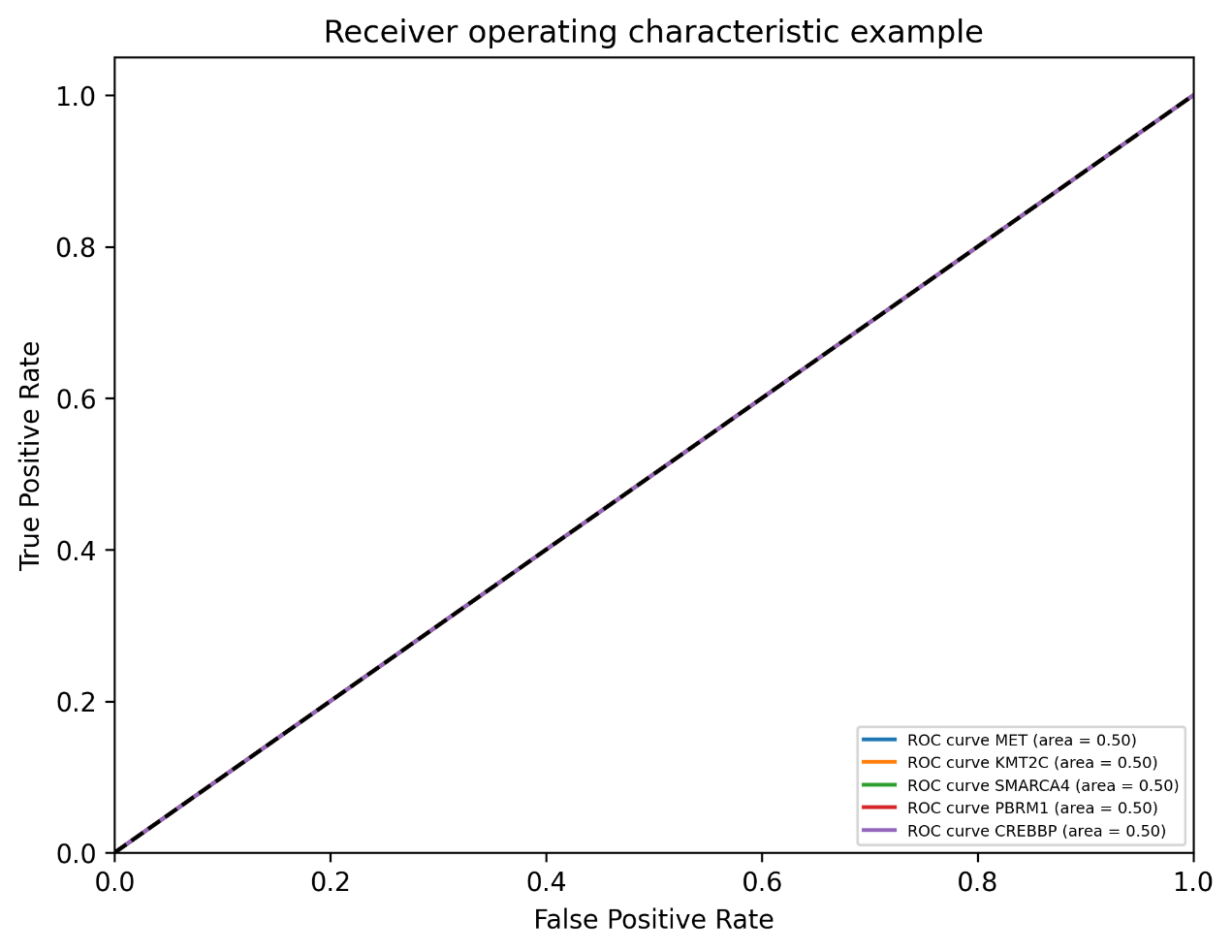
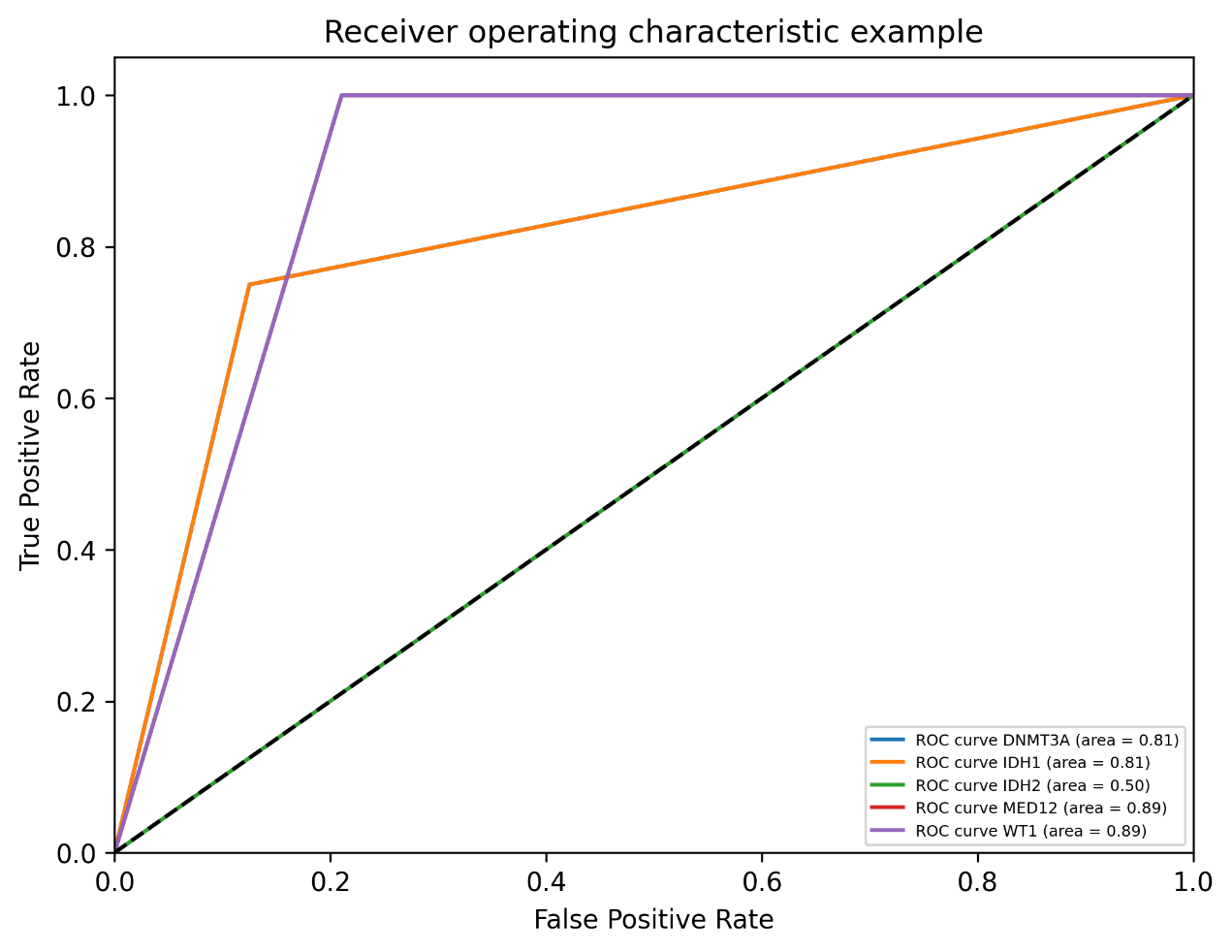
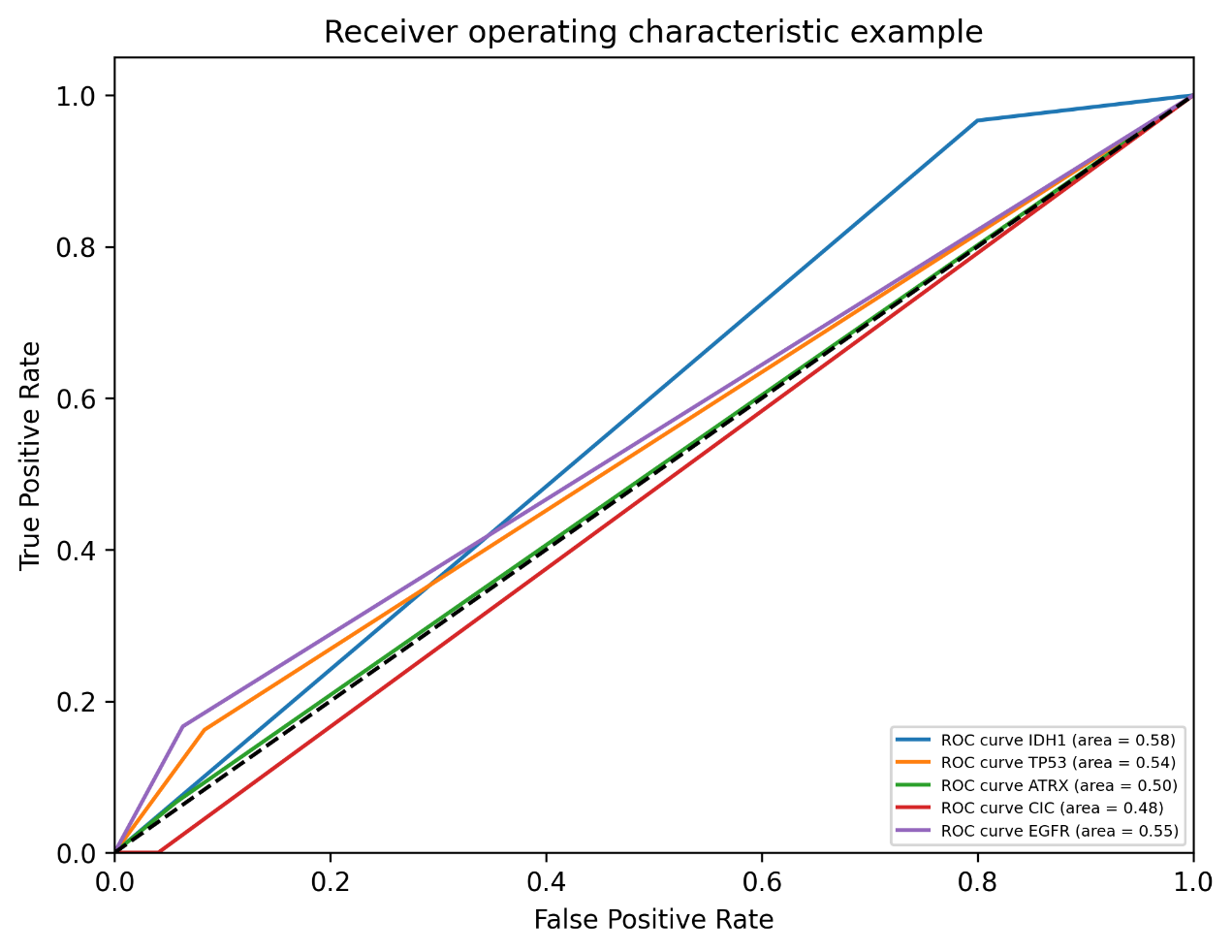
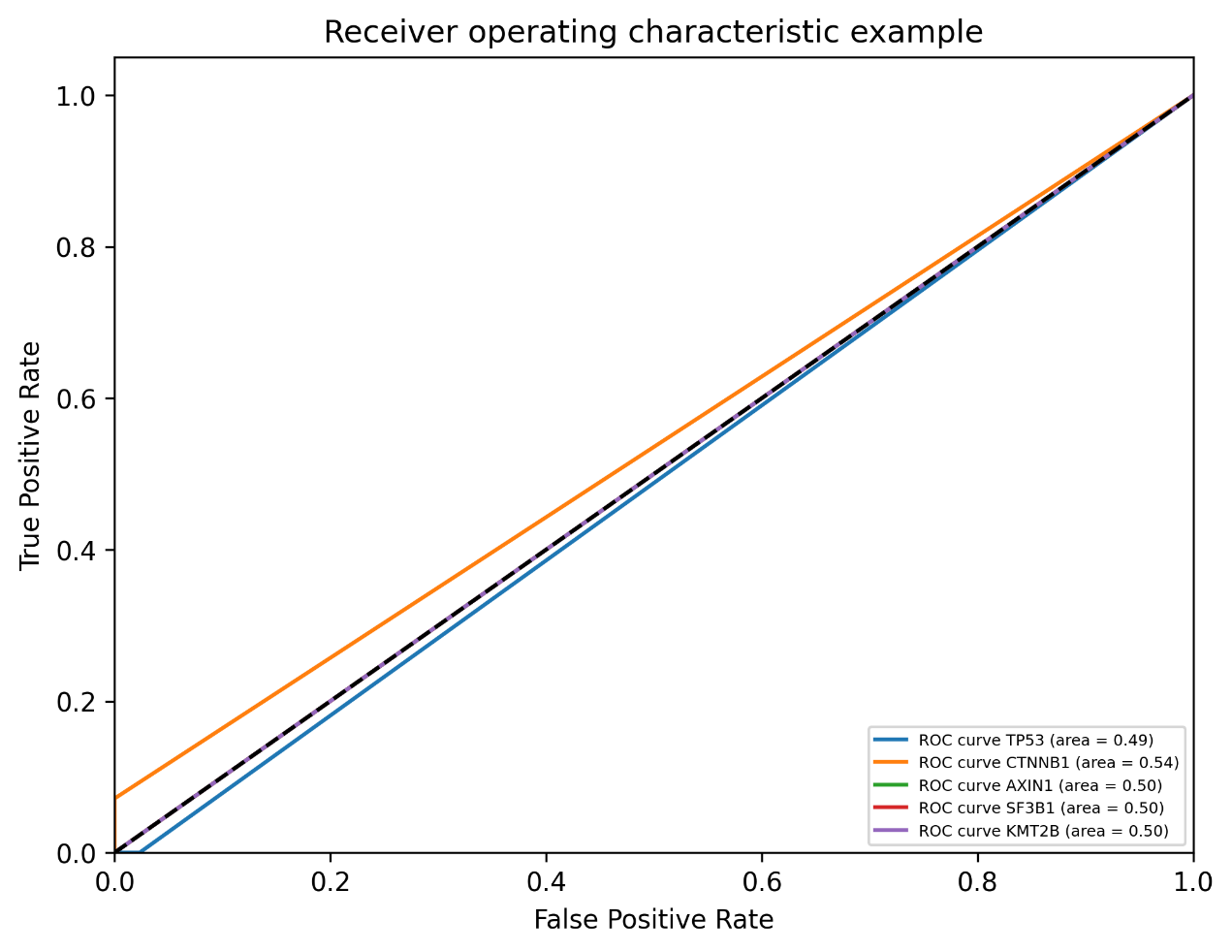
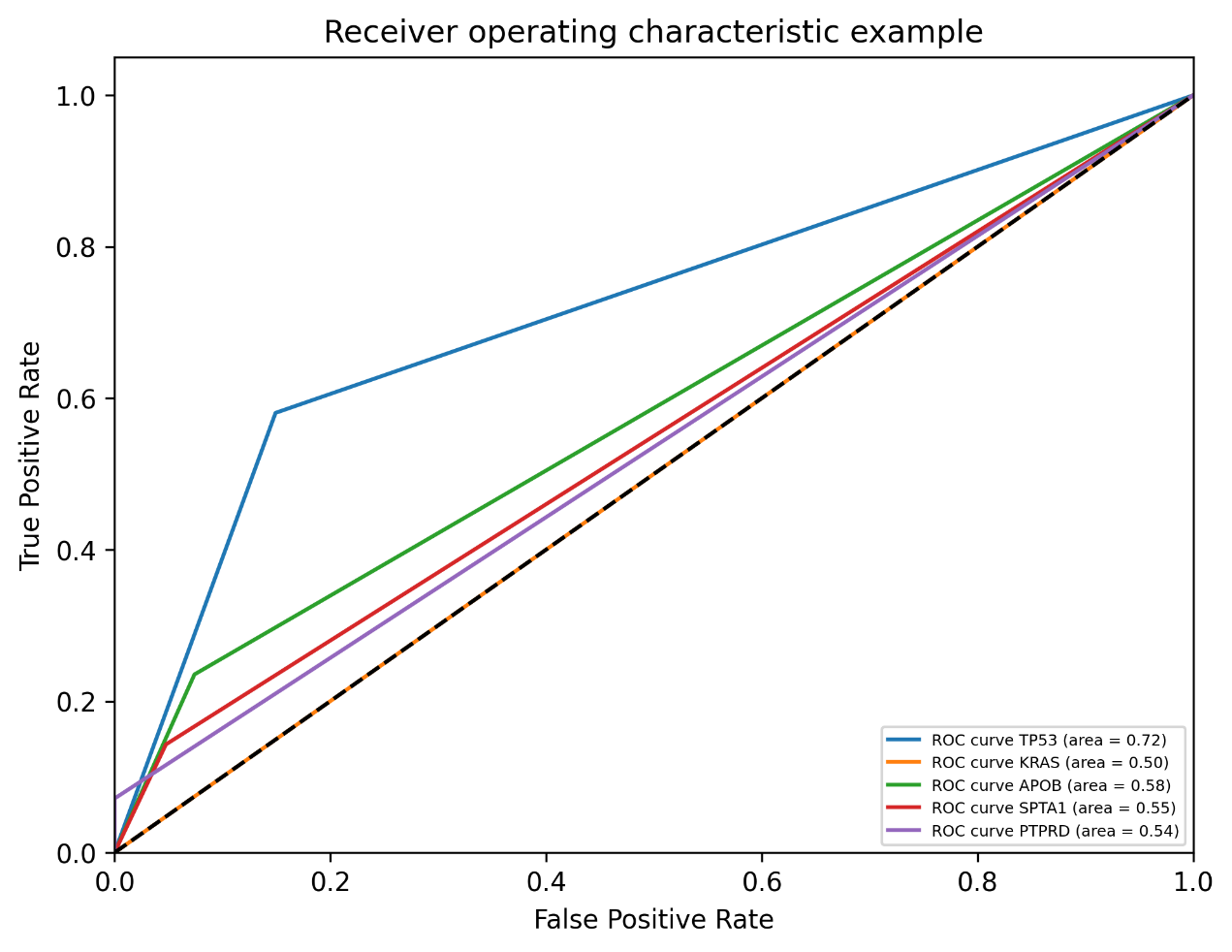
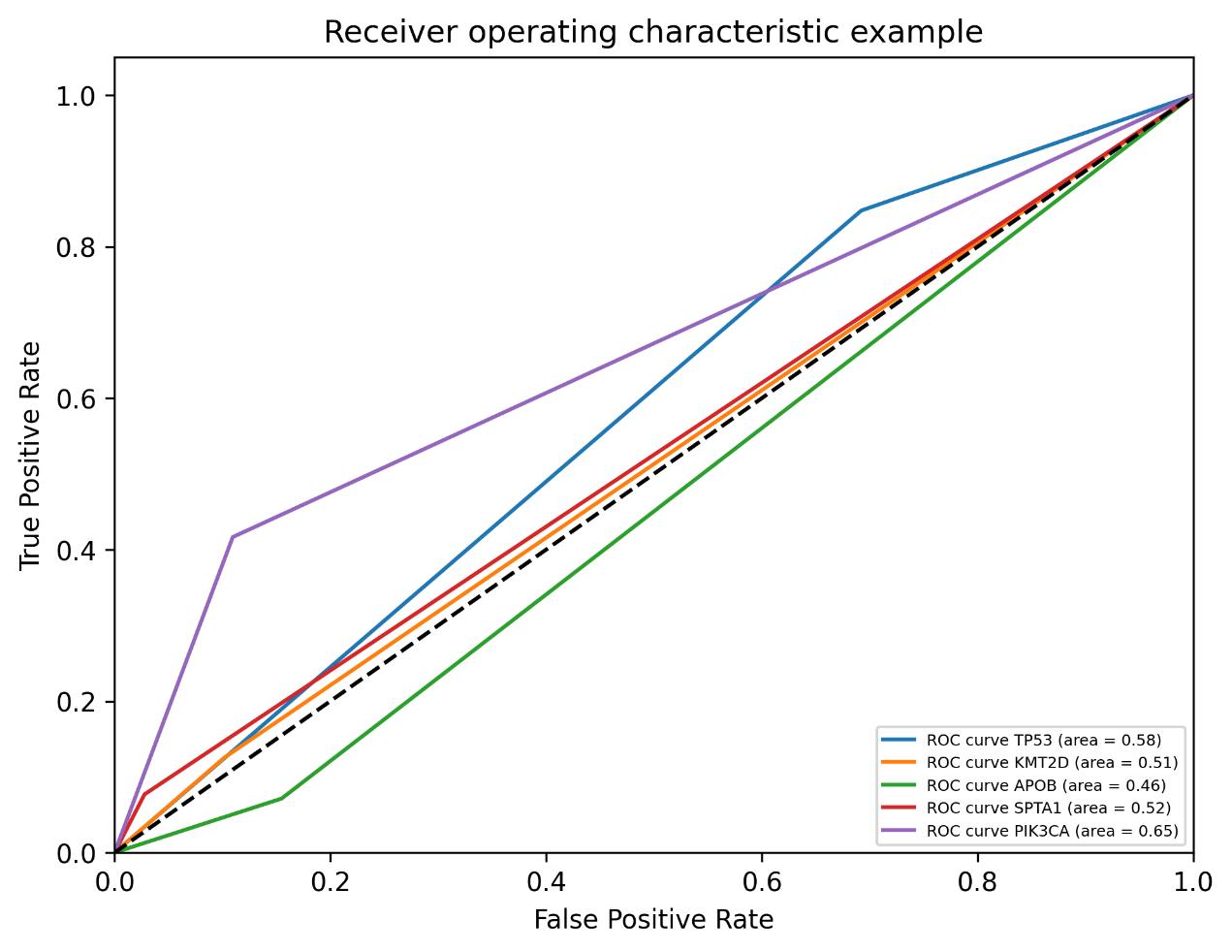
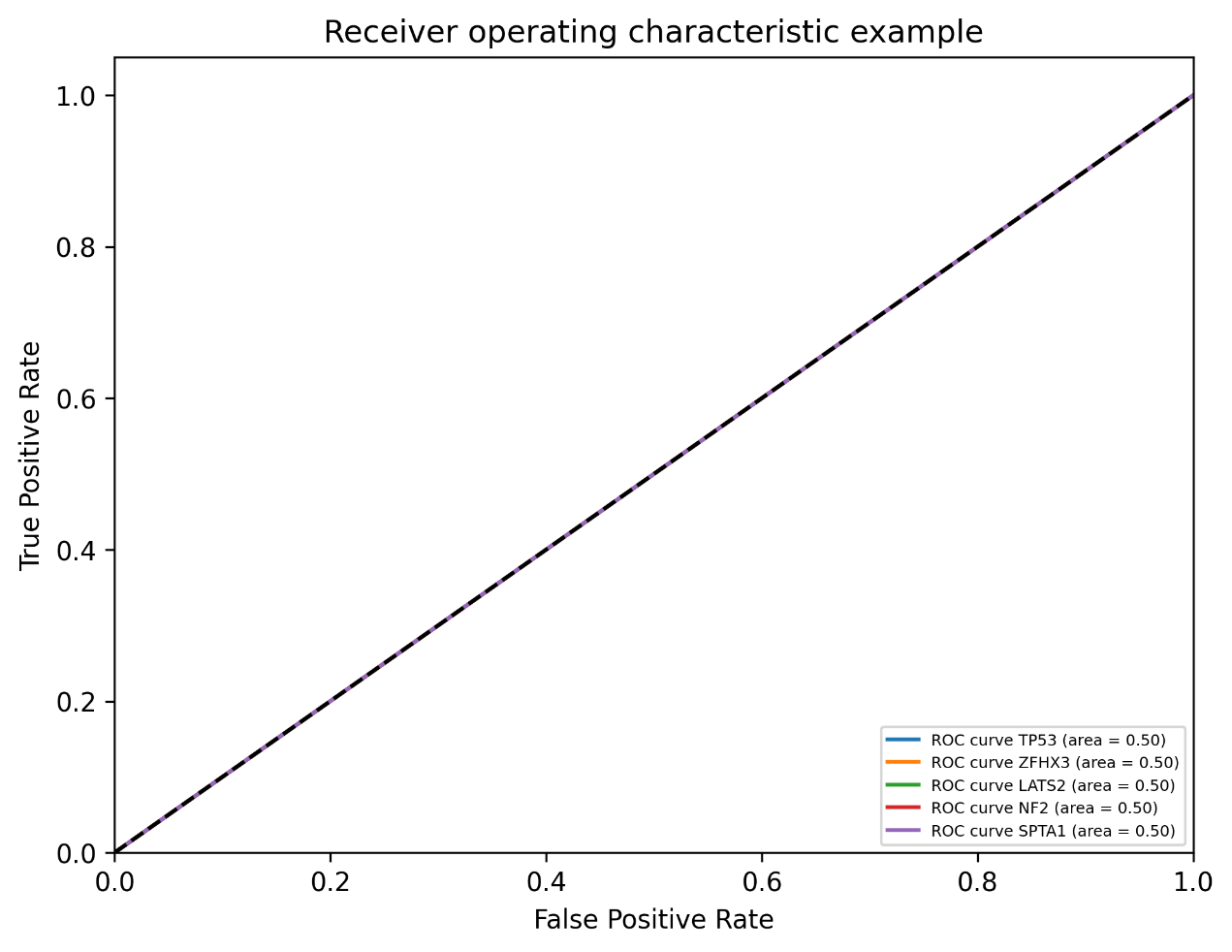
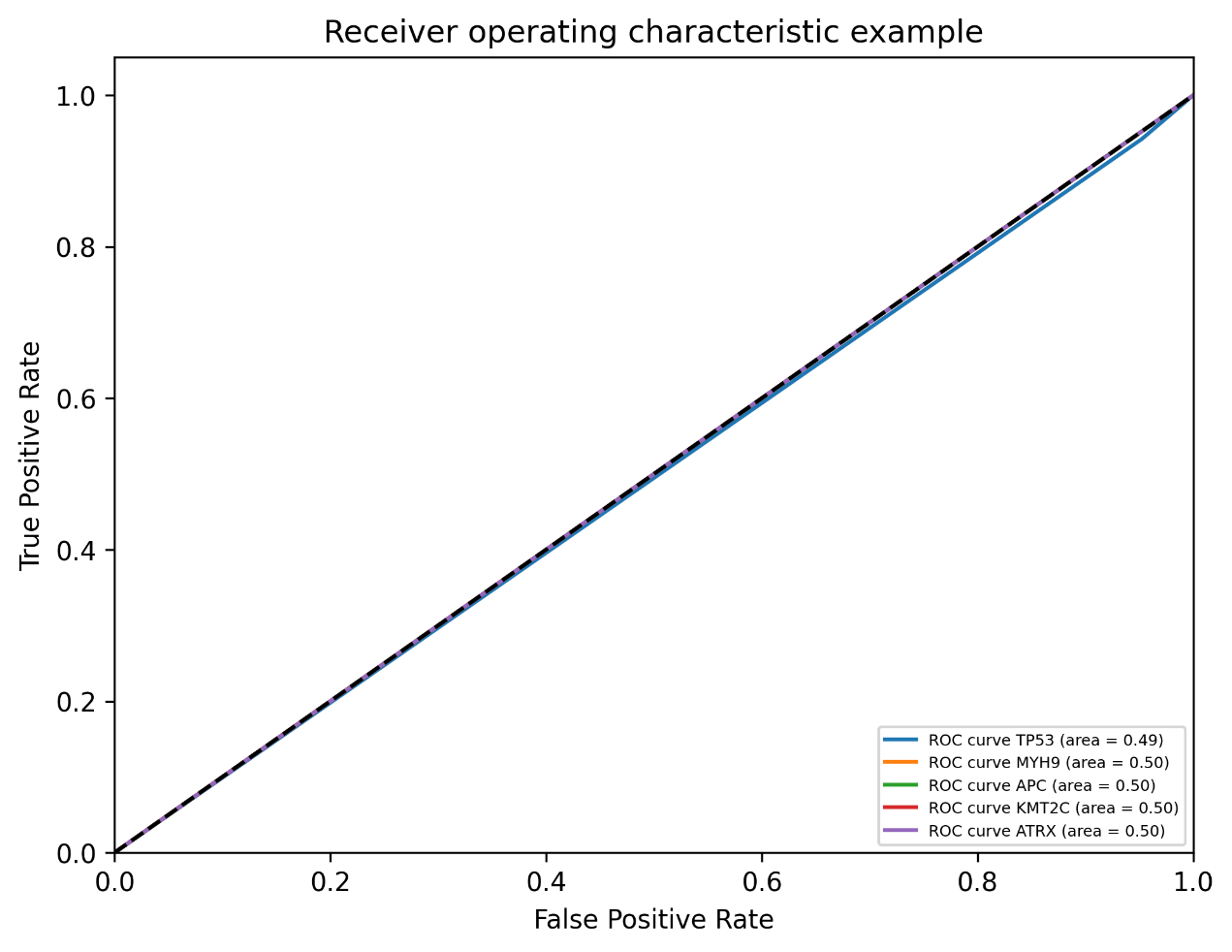
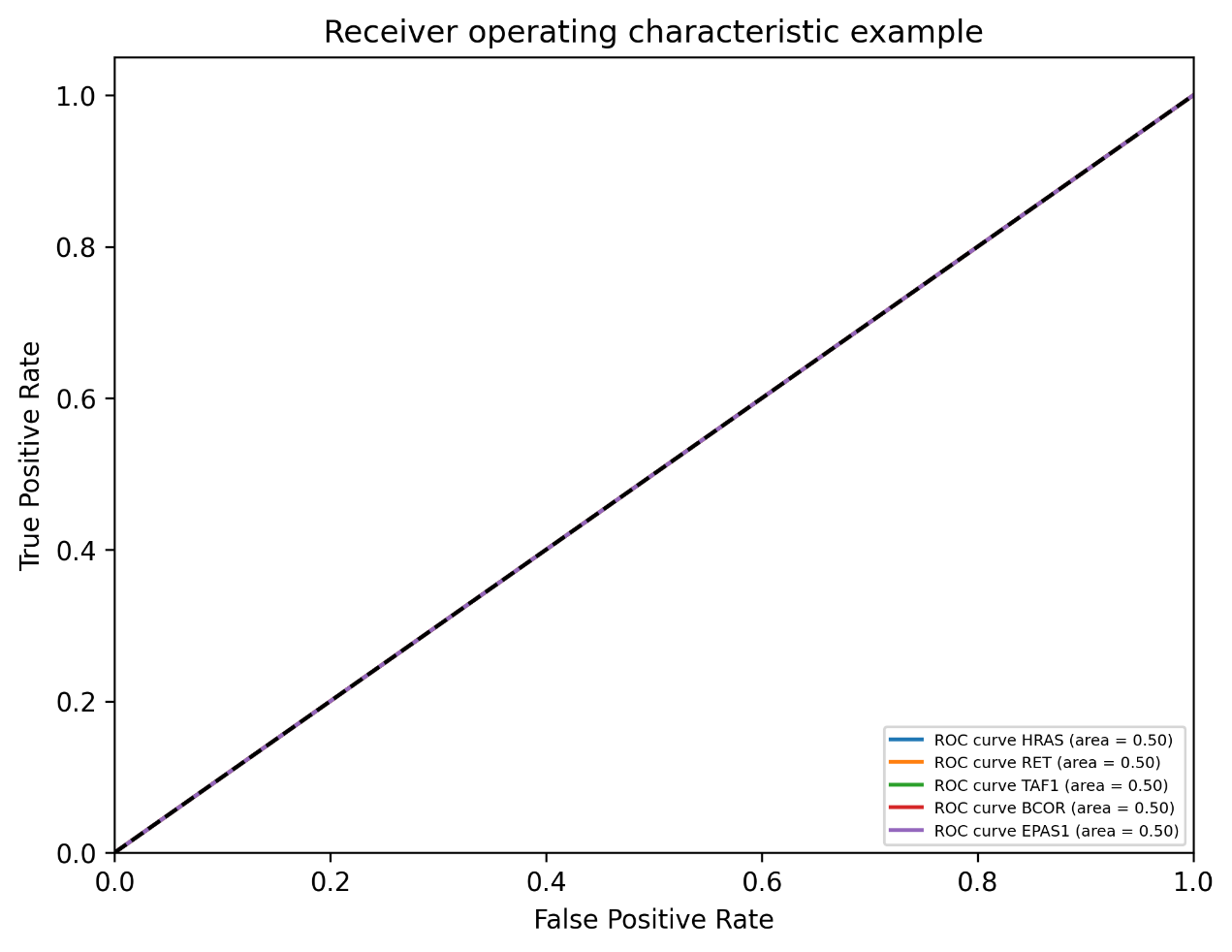
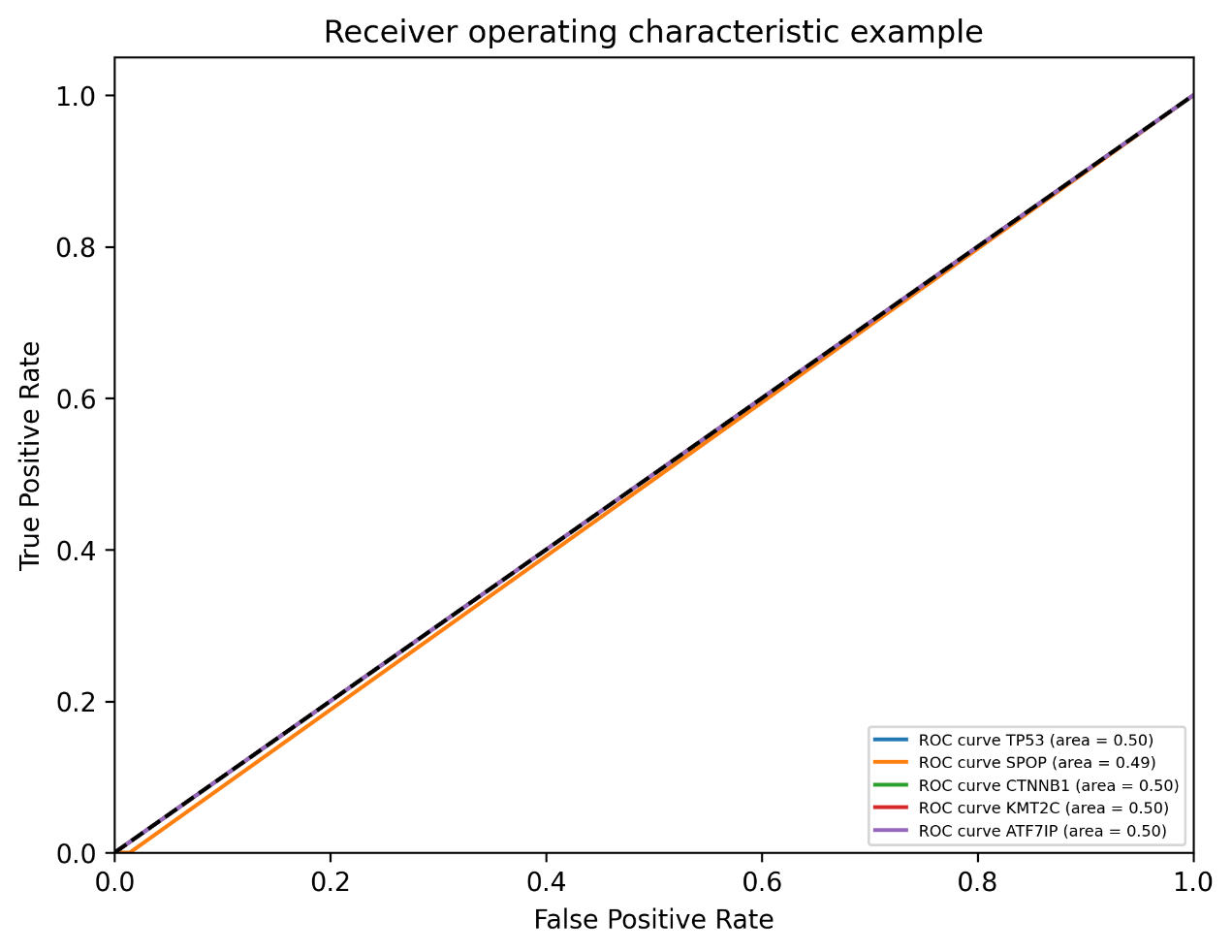
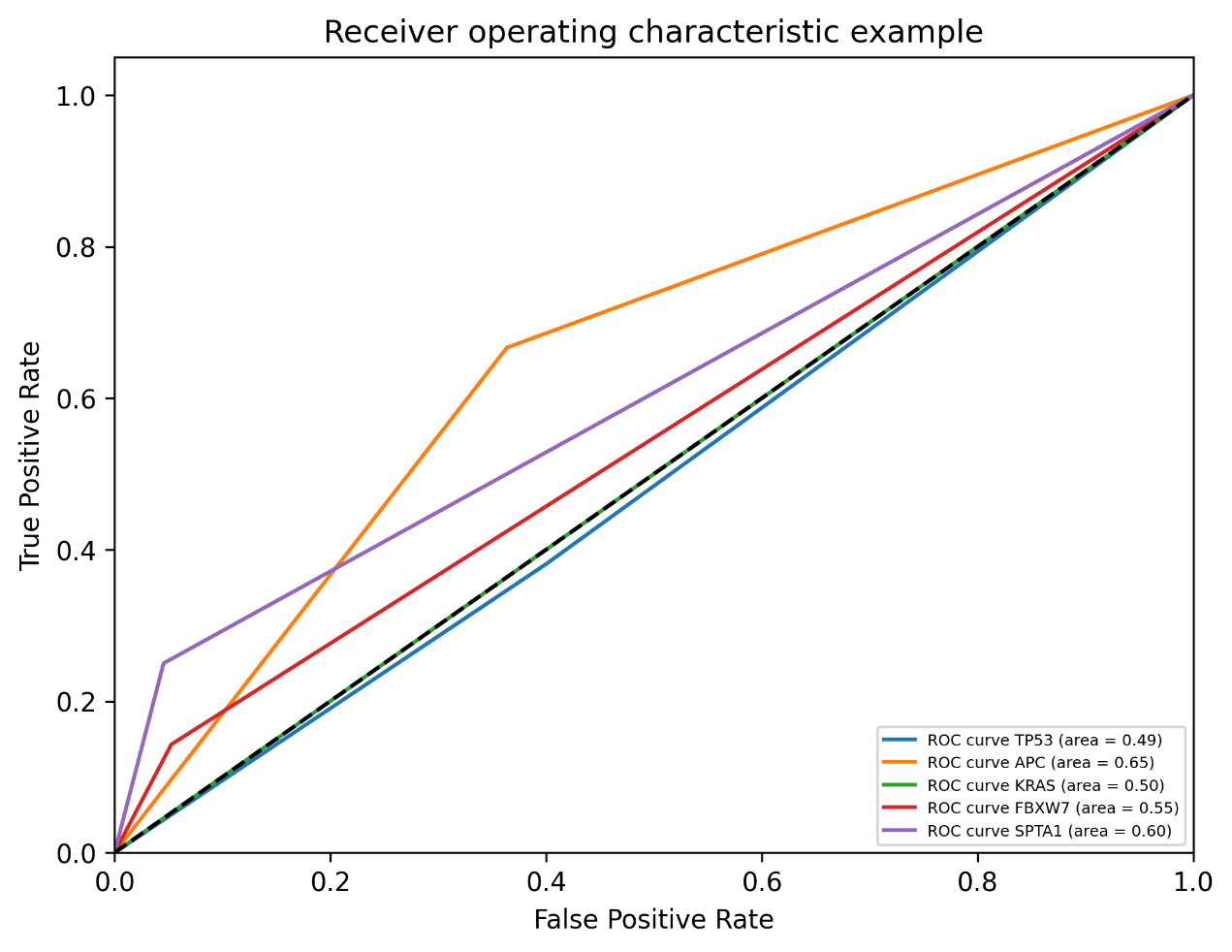
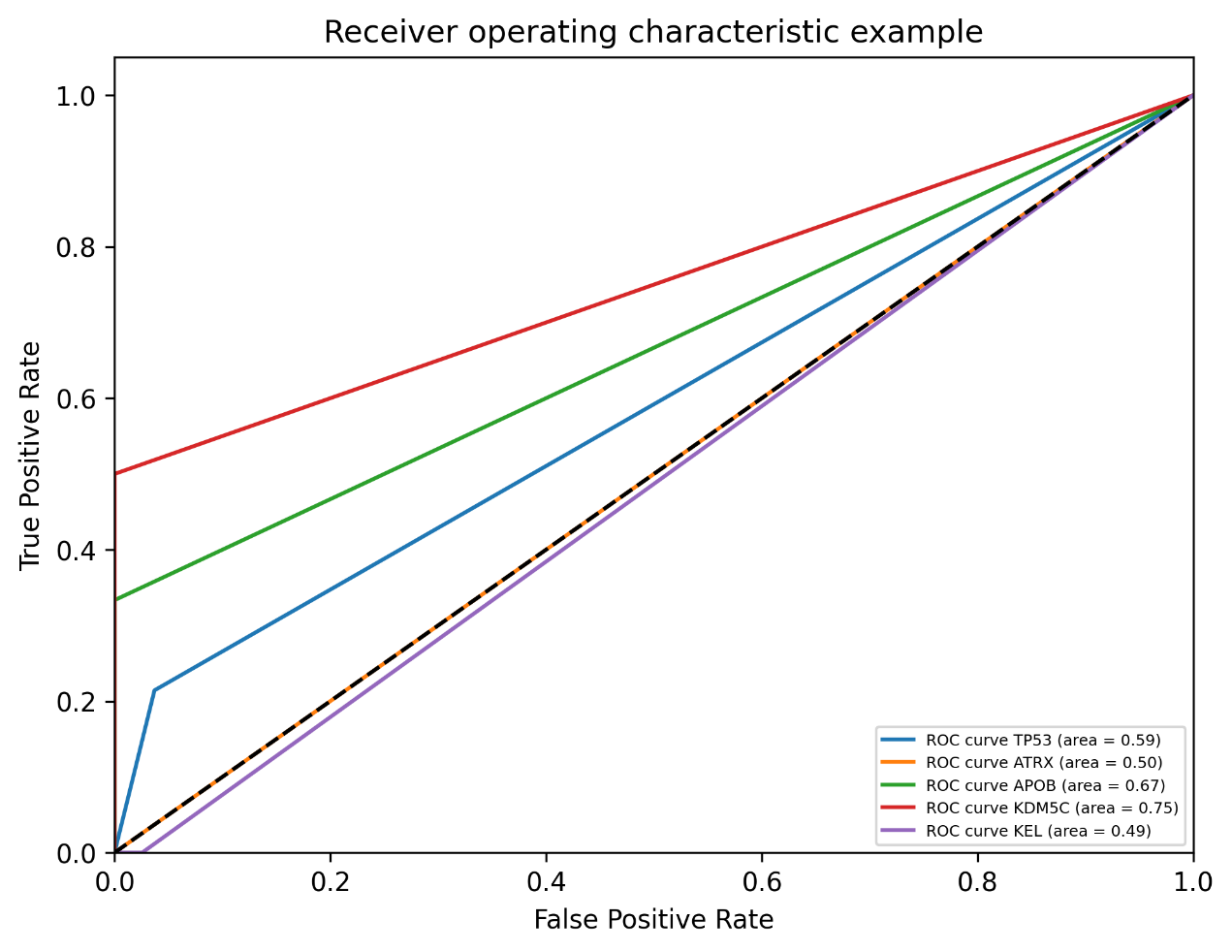
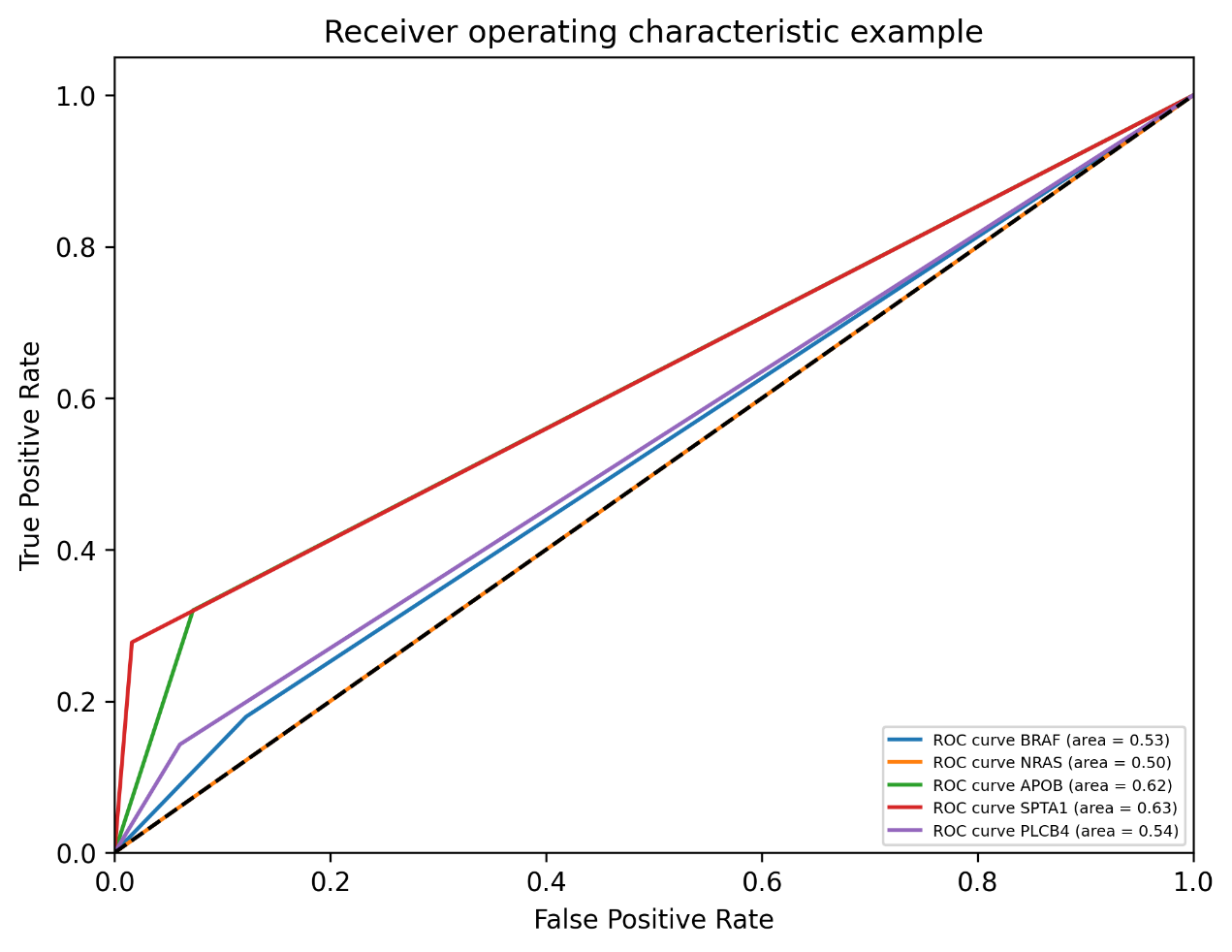
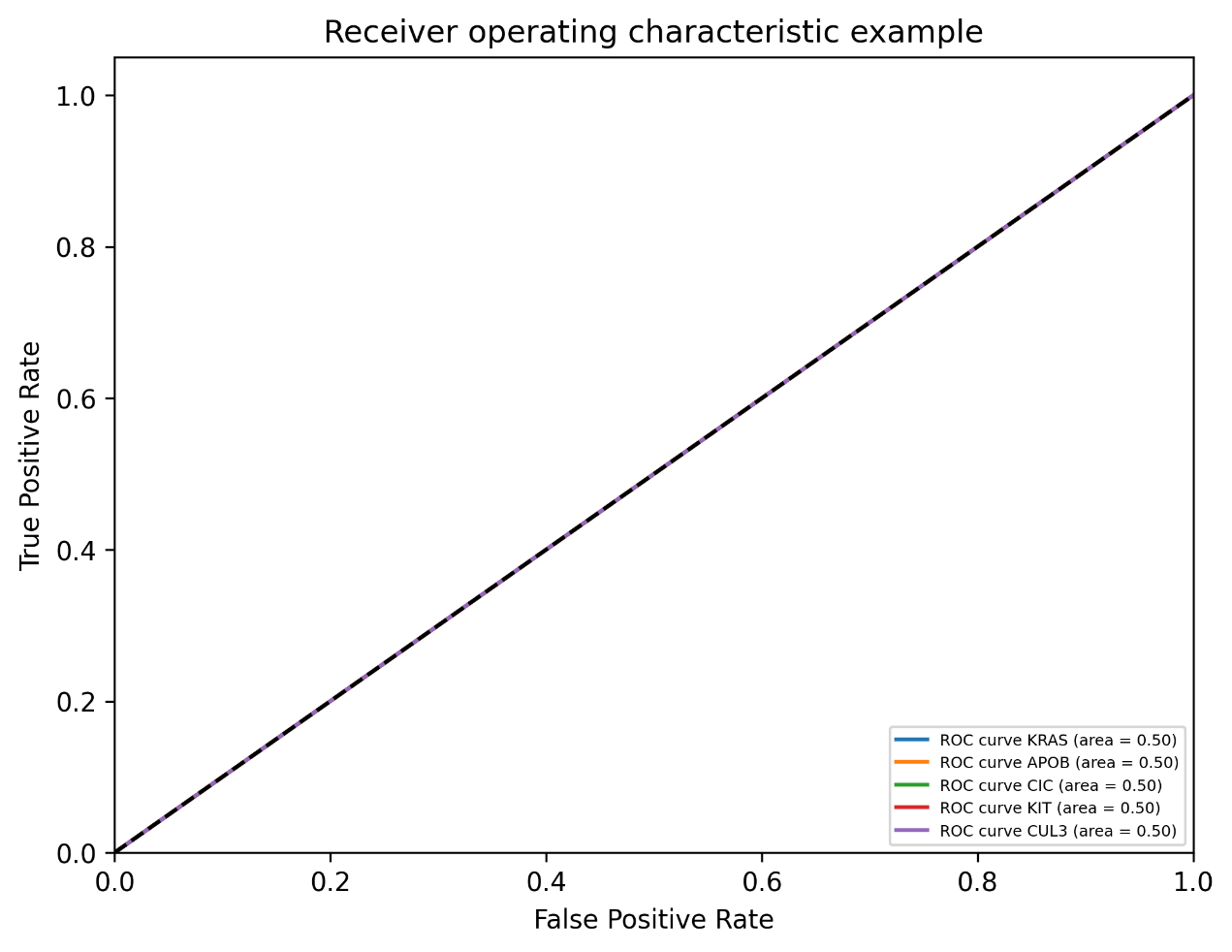
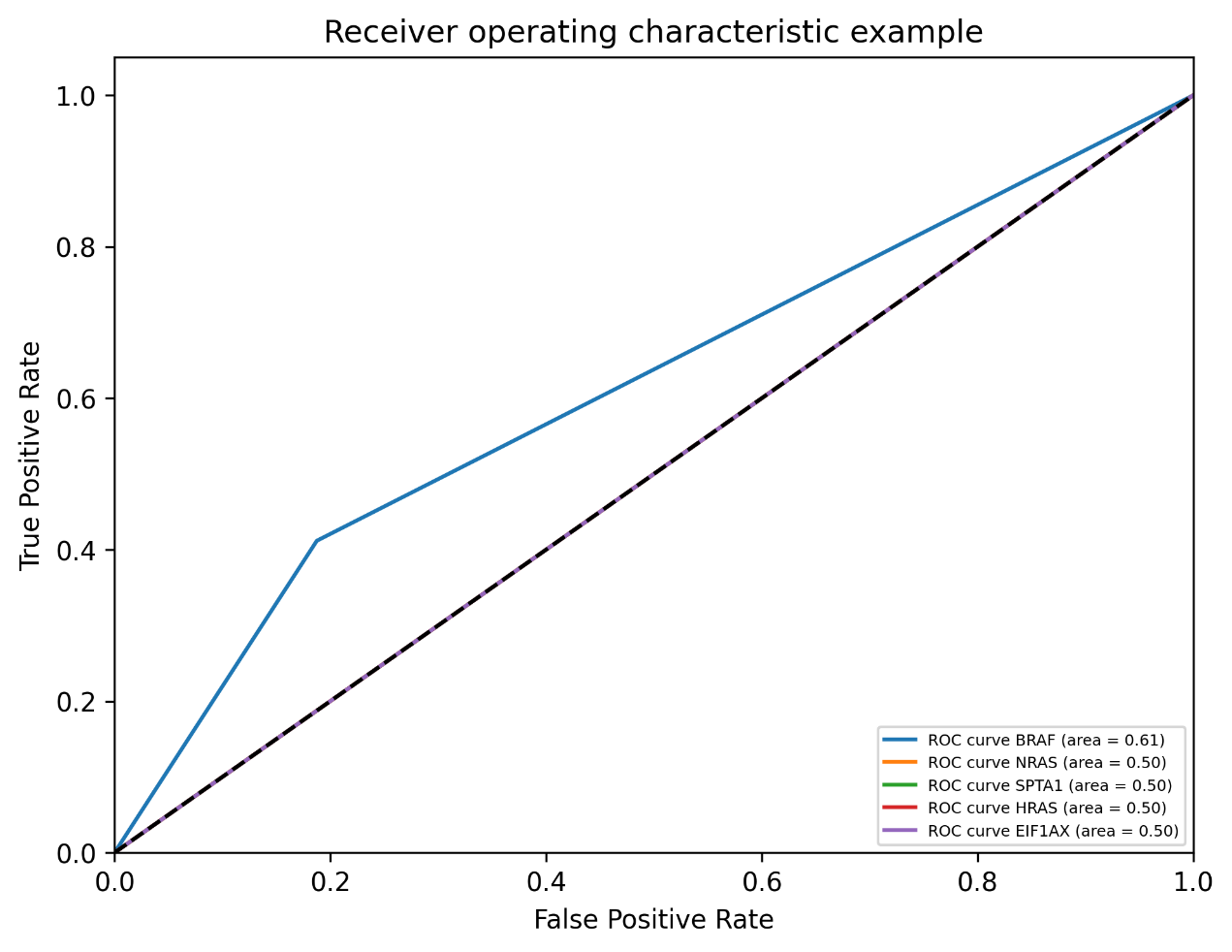
**(2). the result of the classification accuracy of all the fold’s validation dataset for gene is of (5-fold cross validation):**

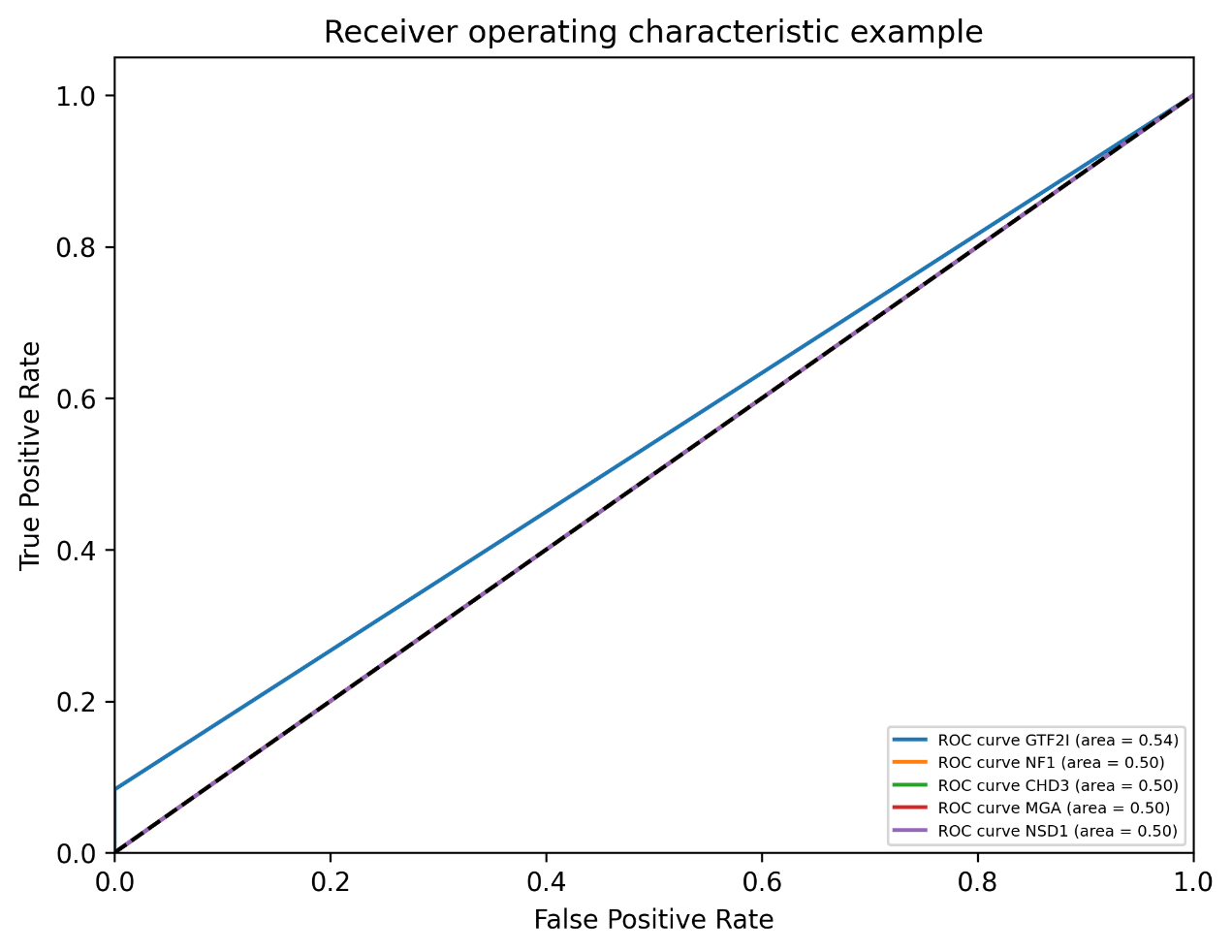
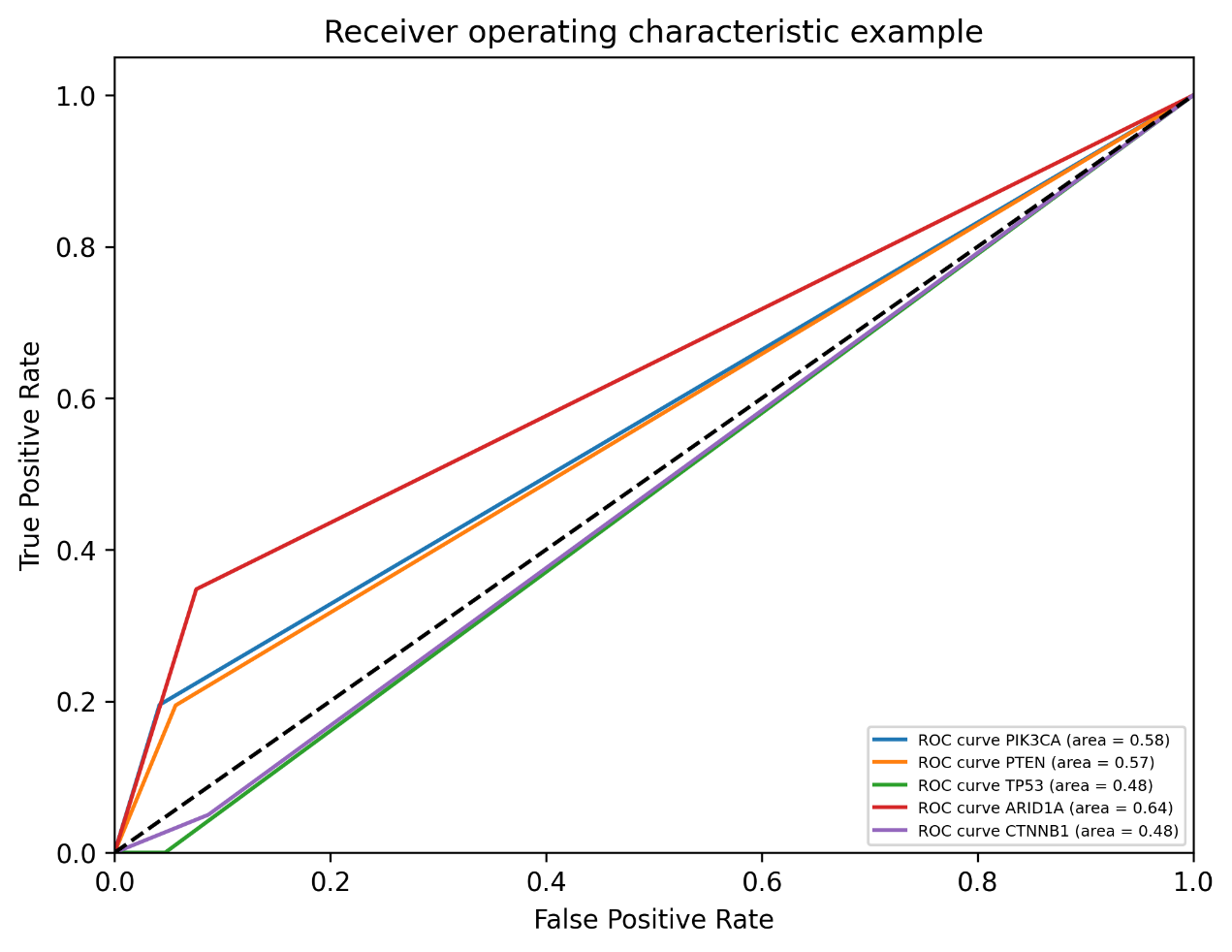
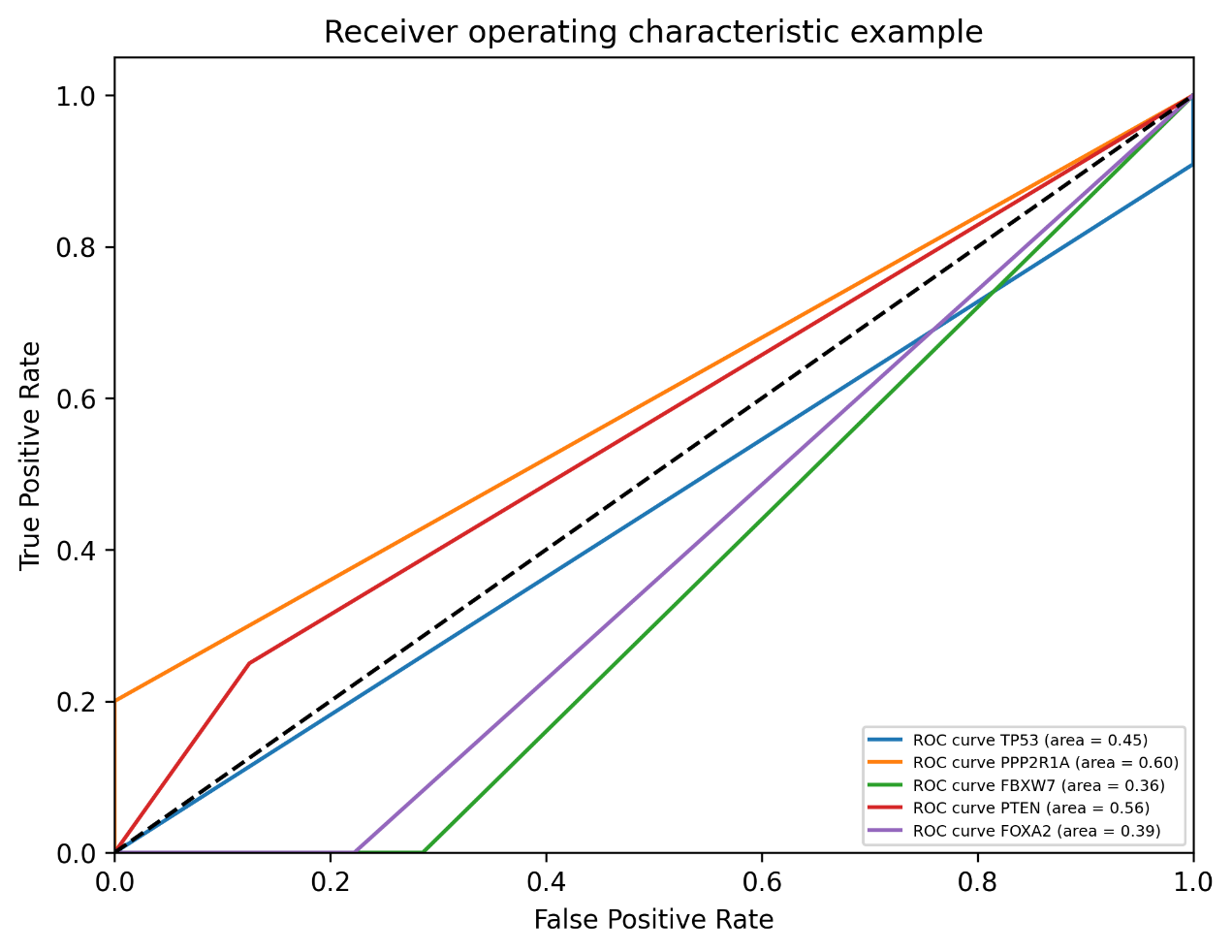
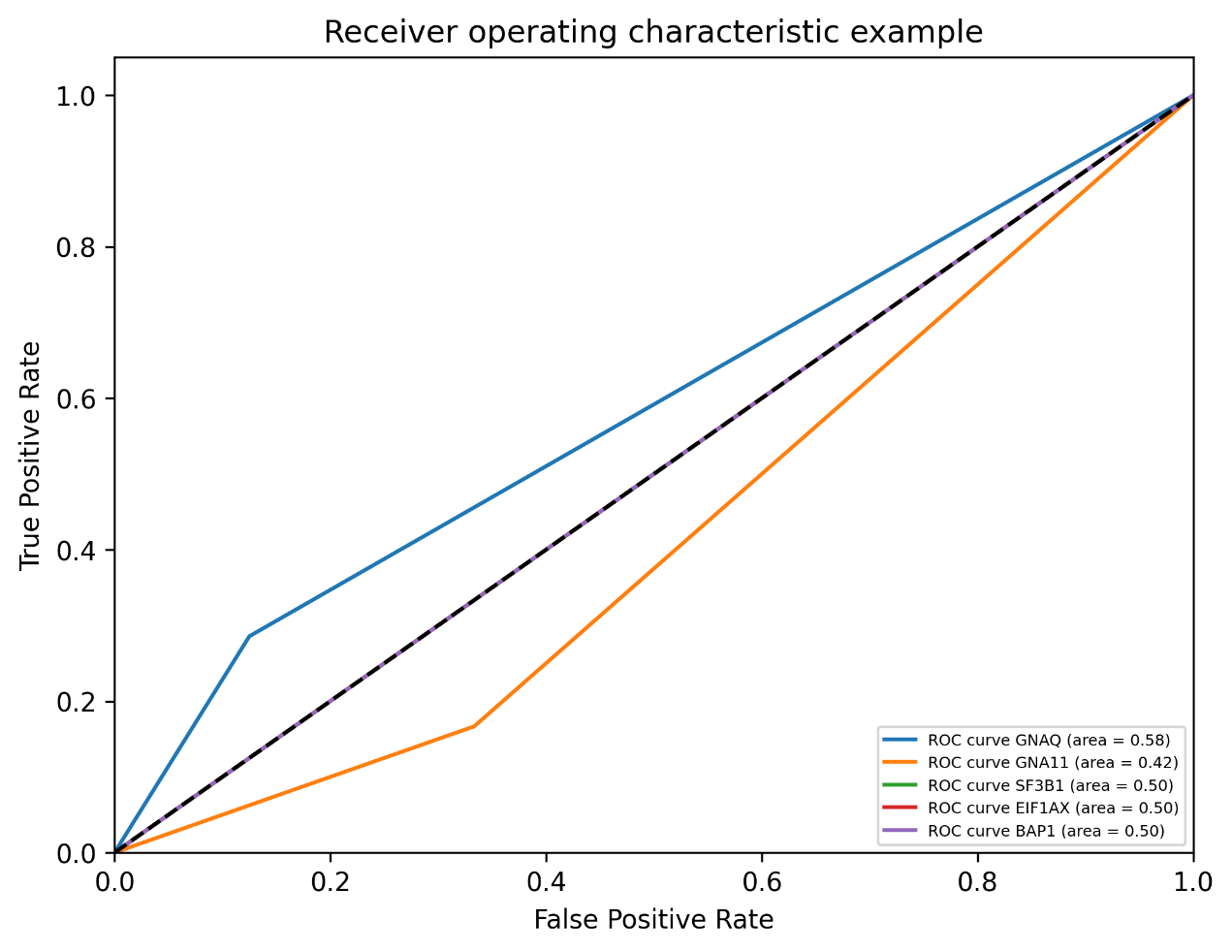
The 5-fold cross validation has 5 testing result, they are:

[0.8374214494515684,0.8375361844139646,0.8395005293757191,**0.8399585230067508**, 0.8391890567606953]

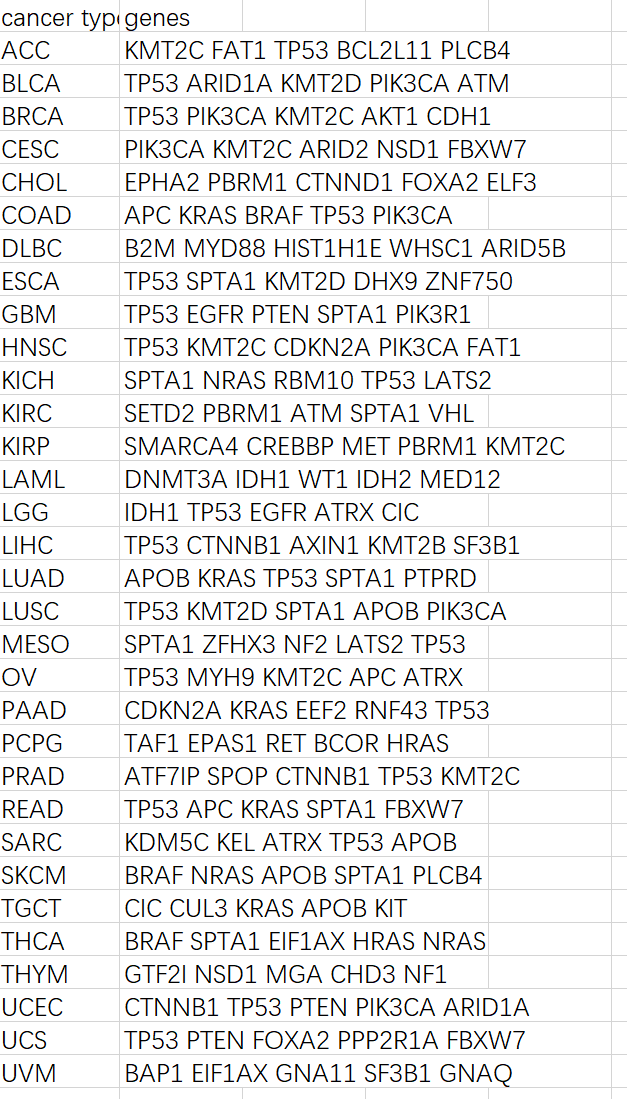
The validation accuracies for 5-fold cross validation are:

[0.7989642706878418, 0.8033426254770113, 0.805973221944154, 0.8062901020228543, **0.8064621150722241**]





**(3).** Find the most determinable genes in each cancer types



This is for the weight of the sbs signature in each cancer types.



1. **The problems:**
2. Can I explain the reason why we classify cancer is that we can use the sbs weight we obtained in each cancer to perform feature extraction to prove that we can use those weighted sbs weight to better predict the top 5 most frequently mutated driver gene in each cancer?
3. The gene classification is not good, how do I mention it in dissertation?

**3.The plan**

(3). working on the dissertation