Global Pan-cancer serum miRNA classifier across 13 cancer types: Analysis of 46,349 clinical samples

Pandikannan Krishnamoorthy [a], Madhavan Parthasarathy [a], Nilanjana Das [a], Athira S Raj [a], Ashok Kumar [b], Vikas Gupta [c], Saikat Das [d], Himanshu Kumar [a, e*]

Affiliations:

[a] Department of Biological Sciences, Laboratory of Immunology and Infectious Disease Biology, Indian Institute of Science Education and Research (IISER) Bhopal, Bhopal-462066, MP, India

[b] Department of Biochemistry, All India Institute of Medical Sciences (AIIMS), Bhopal, Madhya Pradesh, India

[c] Department of Otorhinolaryngology (ENT) – Head & Neck Surgery, All India Institute of Medical Sciences (AIIMS), Bhopal, Madhya Pradesh, India

[d] Department of Radiation Oncology, All India Institute of Medical Sciences (AIIMS), Bhopal, Madhya Pradesh, India

[e] Laboratory of Host Defense, WPI Immunology, Frontier Research Centre, Osaka University, Osaka 5650871, Japan

***Corresponding author:** H Kumar, Department of Biological Sciences, Laboratory of Immunology and Infectious Disease Biology, Indian Institute of Science Education and Research (IISER) Bhopal, AB-3, Room No. 220, Bhopal By-pass Road, Bhauri, Bhopal 462066, MP, India. Tel: +91 755 6691413; Fax: +91 755 669 2392; E-mail: hkumar@iiserb.ac.in

Running head: **Serum Pan-cancer miRNA classifier for detection of cancer**

ABSTRACT

Liquid biopsy offers the minimally-invasive way of early cancer diagnosis. MicroRNAs (miRNAs) are small non-coding RNAs that show promising diagnostic potential due to their stability and their dysregulation upon different physiological conditions. However, existing cancer classifiers often rely on cohort-based comparisons, limiting their clinical utility. Extensive analyses in this study present a pan-cancer miRNA-based single-sample classifier, trained on 16,190 samples, tested across 9 independent datasets, and further validated on 8 distinct disease cohorts. The classifier leverages miRNA expression signatures to classify cancer and non_cancer samples including healthy, other diseases with high sensitivity and specificity, enabling personalized predictions. The classifier identifies cancer by evaluating the relative expression patterns of specific miRNAs, capturing neoplasm-specific dysregulation patterns independent of cohort effects. This study highlights the potential of miRNAs in robust cancer classification, offering a minimally invasive, scalable, and clinically adaptable miRNA serum resource for early cancer detection across diverse populations and malignancies.

Keywords: microRNAs (miRNAs) biomarkers; Single-Sample Classifier (SSC); Pan-Cancer Detection; Liquid Biopsy; Machine Learning Classifier; Early Cancer Diagnosis.

Main

Globally cancer remains a major health burden with nearly 20 million new cases and causing 9.7 million recorded deaths worldwide[1]. Nearly one in every six deaths recorded globally are caused by cancer and the new cancer incidences are predicted to increase up to 35 million by 2025[1,2]. However, overall cancer mortality has decreased to nearly 33% since 1991, due to the advancement in the early diagnosis, treatment methods, population-level screening[3,4]. Therefore, early diagnosis of cancer is crucial.

There are major challenges in the existing early diagnosis methods as they are invasive and cancer tissue-specific. Early diagnosis methods utilizing minimally-invasive liquid biopsy methods involving the clinical samples such as plasma, serum, and non-invasive liquid biopsy method using patient urine and saliva can play an important role in risk stratification. Majority of existing liquid biopsy methods utilize the cancer specific mutations, methylation patterns that have its own limitations such as prevalence of false-positive results due to the presence of other diseases[5].

The MicroRNAs (miRNAs) are small non-coding RNAs that are readily present in the extracellular fluids and show differential patterns associated with clinical symptoms and with detectable tumors. [6,7]. The miRNAs are highly stable in the blood serum and possess diagnostic potential. Several studies utilized advanced statistical methods, machine learning algorithms to identify miRNA-based classifiers in blood components such as serum and plasma. However, the biggest drawback limiting their clinical applications is that miRNAs are dysregulated in several diseases and the cancer-specific signature often overlaps with any other diseases associated with

inflammation. The existing databases such as The Cancer Genome Atlas (TCGA) only consist of cancer tissue-specific omics datasets and the cancer miRNA profiles derived from the serum samples are not harmonized in a single database. There are very limited resources for the liquid biopsy transcriptome in larger sample sizes to infer any models with high confidence.

By employing the rules-based binary classification algorithms, this study identified the rules-based classifier model utilizing the set of four rules, discriminate between cancer and non_cancer samples including healthy individuals and patients from other non-infectious diseases with highest accuracy. The accuracy of the model was also validated in the independent datasets comprising the cancer and other diseases.

Results

Mining of Serum miRNA datasets across cancer types:

TCGA lacks the miRNA expression datasets derived from blood serum and hence the public transcriptome datasets such as GEO (Gene Expression Omnibus), Array Express were queried for the eligible miRNA expression datasets derived from serum. A total of 46349 samples from the independent datasets were identified and screened for the analysis (**Supp. Fig. 1**). 490 samples were removed due to the presence of duplicates across datasets and improper labelling. The distribution of samples across each cancer type in each dataset were visualized in (**Figure 1A).** The dataset GSE211692 was utilized as the training dataset due to its higher sample size of 16190, out of which 9921 samples belong to the "cancer" class and 6269 samples belong to "non_cancer" class [8]. The datasets GSE122497[9], GSE106817[10,11], GSE137140[12], GSE113740[13], GSE112264[14], GSE139031[15], GSE113486[16], GSE164714[17] and GSE73002[18] were utilized as test datasets that comprise of both "cancer" and "non_cancer" class for which the distribution is visualized in (**Fig.1 A)** (**Supp. Table 1**). The datasets GSE110317[19], GSE110651[20], GSE134108[21], GSE85589[22], GSE117064[23], GSE120584[24], GSE150693[25]and GSE140249[26] were considered validation datasets that comprises of either "cancer" or "non_cancer" class. Even if both classes were present, at least one class had very less sample size. The "non_cancer" class consist of diverse non-cancer diseases such as Autoimmune Hepatitis, Alzheimer's disease (AD), Vascular Dementia (VD), Mild cognitive Impairment (MCI). The purpose of validation datasets is to test whether the classifier is able to correctly predict "non_cancer" samples with identified rules. Overall, 46349 serum samples profiled for miRNA expression were included in the study.

Out of which 16190 samples, 25082 samples and 5077 samples were utilized as training dataset, test datasets and validation datasets, respectively.

The conventional machine learning classifier analyzes overall miRNA (features) expression variation across the samples that belong to different class(es) and identifies discriminatory features through statistical or machine learning algorithms. The top discriminatory features, comprising significantly dysregulation between each class, will be utilized to build the classifier model. The performance evaluation of this classifier can be affected by class imbalance, batch effects between independent datasets, requirement of appropriate normalization techniques and more importantly test and validation datasets require the same number of classes as in training datasets. Rules-based single sample classifier utilized in this study identifies pairwise comparisons of each miRNA with other miRNAs and establishes the relationship within a single sample. The rules-based classifier in the format of expression of miRNA A greater than miRNA B denotes “cancer” class and reverse is “non_cancer” class. The rules-based classifier obtained through this approach can be validated in independent datasets and their performance will not be affected by batch effects or normalization methods and also the validation datasets need not have the same number of classes. For clinical diagnosis and screening in a large population, this approach can be valuable. The differences between our approach and the conventional approach is depicted in (**Fig. 1B).**

Top Serum miRNA rules that can discriminate between “cancer” and “non_cancer” class:

The training dataset GSE211692 comprises samples belonging to 13 different cancer types and non_cancer samples along with benign samples. Distribution of cancer types in the dataset is depicted through Lollipop plot (**Fig. 2A**). The expression of 2565 miRNAs across 16190 samples

were subjected to filtering in which miRNAs with minimal expression in more than 10% of the samples were discarded. 211 miRNAs remained for the further analysis (**Fig. 2B**). The top scoring pairs (TSP) were identified for each sample per class and the top rules identified through each class were identified. The segregation between the "cancer" and the "non_cancer" class was visualized through the t-SNE plot (**Fig. 2C)**. A total of 4 TSPs were identified to segregate the "cancer" and "non_cancer" classes based on the scores that depicts how closely the pairs are associated with the class (**Fig. 2D**). miR-4258 > miR-4730, miR-663a > miR1228-5p, miR-6787-5p > miR-6800-5p and miR-1469 > miR-1268a are the top rules that are specific for "cancer" class and confusion matrices denotes the relationship between actual and predicted labels for the rules (**Fig. 2E – 2H**). Overall, the classification is best for the first two rules – miR-4258 > miR-4730 and miR-663a > miR-1228-5p, followed by other two rules of the training data.

Construction of Serum pan-cancer single sample classifier using four rules:

The expression of the eight miRNAs forming four rules were visualized through beeswarm plot (**Fig. 3A**). There is a clear distinction between the expression pattern of the miRNA pairs in "cancer" and "non_cancer" class. Notably, expression level of each miRNAs is higher and hence it has higher diagnostic potential. The overall performance metrics of each of the rules are visualized using heatmap (**Fig. 3B**). The rules 4730>4258 and 663a>1228_5p performs better than the other two rules individually. Relying on any one rules can affect the specificity across diverse disease conditions and hence the classifier model was built based on the vote-based approach where more than two rules have to be satisfied to be predicted as "cancer", and

predicted as “non_cancer” when two or less than two rules are satisfied (**Fig. 3C**). The pan-cancer classifier with this voting-based approach was evaluated in the training dataset and the classifier performance improved as a whole as depicted in the confusion matrix (**Fig. 3D**). The confusion matrix depicts that 9897 samples out of 9921 samples were correctly predicted as “cancer”, whereas 5626 samples out of 6269 samples were correctly predicted as “non_cancer”. Only 24 “cancer” samples were wrongly predicted as “non_cancer”, whereas 643 “non_cancer” samples were predicted as “cancer”. The heatmap denotes correct predictions (light green) and incorrect predictions (light red) across 16190 samples of training data (**Fig. 3E**) (**Supp. Table 2**). The cancer types among the cancer class are also highlighted. Overall, the 4-rules based serum pan-cancer classifier predicts the “cancer” and “non_cancer” with higher accuracy while the performance of classifier is affected while predicting certain “non_cancer” samples that are predicted as “cancer”.

Classifier performance evaluation in independent test datasets:

The performance of the classifier is tested in 9 independent test datasets. Test datasets comprise of both “cancer” and “non_cancer” classes samples derived from individual studies that are not part of the training dataset. The overall distribution of disease state of both the “cancer” and “non_cancer" class of samples across all the 9 datasets were visualized through the tile plot (**Fig. 4A**). The distribution of samples between “cancer” and “non_cancer” class were visualized through t-SNE plots for all the 9 test datasets (**Supp. Fig. 2 A-I**). The expression levels of eight miRNAs that are the part of pan-cancer classifier was plotted and visualized through beeswarm plots for all the test datasets that clearly tells these miRNAs are having higher expression and

also the change in the expression patterns between two classes were distinct (**Supp. Fig. 3 A-I**). The classification of individual rules was plotted through the confusion matrices for all test datasets (**Supp. Fig. 4 B-J**). Individual rules experience varied classification results. The status of predictions by the pan-cancer classifier model was visualized through the confusion matrices (**Fig. 4B**). As observed in the training datasets, the classifier performance is good in predicting "cancer" samples as "cancer" accurately in almost all the datasets. The performance of the classifier was poor while predicting few "non_cancer" samples as the classifier misclassify them as "cancer". The error is more in GSE112264 dataset as 242 samples out of 262 "non_cancer" samples were predicted as "cancer". However, all the 1309 cancer samples were predicted as "cancer". The heatmap denotes the prediction results of all the 25082 samples present in nine test datasets (**Fig. 4C**) and details of the predictions of all samples are provided in (**Supp. Table 3**). The disease state distribution for the "cancer" and "non_cancer" class were also observed through this plot. The heatmap of performance metrices of the pan-cancer classifier over all test datasets shows poor performance of the GSE112264 dataset in terms of F1 score, Recall and sensitivity due to misclassification of the "non_cancer" samples as "cancer" samples (**Fig. 4D**). The diagnostic performance of the pan-cancer classifier was visualized through area under the Receiver Operating curve (ROC) plots and that shows poor performance in GSE112264 dataset (**Fig. 4E**). Overall, pan-cancer classifier predicts "cancer" samples as "cancer" accurately in 9 independent datasets whereas the performance was poor in prediction of "non_cancer" samples.

Classifier performance in validation datasets comprising other diseases:

The biggest advantage of single sample classifier is that the performance of the classifier can be evaluated even in datasets without requirement of all classes present in the training process. Eight independent validation datasets were identified that has either one of the “cancer” and “non_cancer” classes or imbalanced sample sizes between both the classes. The distribution of the disease state of all the samples across the independent validation datasets were visualized through the tiles plot (**Fig. 5A**). Since the classifier performance was not up to mark in prediction of “non_cancer” class, validation datasets contain the majority of samples from other disease conditions apart the cancer. Some of the other diseases includes Alzheimer’s disease, Autoimmune diseases, Dementia, Mild cognitive impairment, Auto-immune Hepatitis. The predictions of each rules were identified through the confusion matrices (**Supp. Fig. 5 B-I**). The performance of the pan-cancer classifier in the validation datasets was visualized through the confusion matrices (**Fig. 5B**). The dataset GSE110317 comprises of 958 samples, out of which 921 were breast cancer and 37 were benign breast cancer samples. All the 37 benign breast cancer samples were predicted as “cancer” instead of “non_cancer” class. However, 918 out of 921 breast cancer samples were predicted as “cancer”. GSE110651 datasets comprises of 147 metastatic breast cancer samples and 145 of them were correctly predicted as “cancer”. GSE117064 dataset comprises of 1785 cerebrovascular diseases and non-CVD control samples and 1776 samples were correctly predicted as “non_cancer” class. GSE120584 comprises of 1601 samples of neurological disorders such as Alzheimer’s disease, Dementia and non-disease controls. 1596 samples were correctly predicted as “non_cancer” class by the classifier. GSE134108 dataset comprises of 71 samples of both the metastatic and non-metastatic brain cancer. All of the samples were perfectly predicted as “cancer”. GSE140249 dataset comprises of 142 samples of autoimmune hepatitis, Biliary Cholangitis and 133 of these samples are

predicted as “non_cancer” class. GSE150693 dataset comprises of 197 Mild cognitive Impairment patients with or without Alzheimer’s disease conversion and 196 of them were predicted accurately as “non_cancer” class. The GSE85589 dataset comprises of 232 samples of both “cancer” and “non_cancer” samples that includes pancreatic cancer, Cholangiocarcinoma, colorectal cancer and non-cancer controls. 192 of 203 cancer samples were predicted correctly as “cancer”. However, all the 29 non-cancer samples were predicted as “cancer”. The overall prediction results of validation datasets and the disease state distribution was observed through the heatmap (**Supp. Fig. 5J**) and details of the prediction results of each sample is provided in (**Supp. Table 4**). The number of correct (light green) and incorrect predictions (red) in each of validation datasets were visualized through bar plot (**Fig. 5C**). The distribution of disease state for only the incorrect predictions was visualized through the tiles plot (**Fig. 5D**).

Interestingly, the pan-cancer classifier wrongly predicted as cancer in 37 non-cancer samples and all of them were benign breast cancer samples in the dataset GSE110317. All the incorrect predictions of the classifier in test datasets and training datasets were explored to see if there is any pattern among the wrong predictions. The number of incorrect predictions in the test and training datasets were visualized through bar plots (**Fig. 5E)** and (**Fig. 5G**) respectively. The distribution of the disease state of the incorrectly predicted samples in test datasets and training dataset were visualized through the tile plot (**Fig. 5F**) and bar plot (**Fig. 5H**) respectively. In the test datasets, 314 wrong predictions were observed in GSE106817, out of which 306 samples were of benign origin. 29 benign ovary, 62 borderline ovary and 215 benign non-cancer samples were predicted as “cancer”. Similarly, all 242 wrong predictions of GSE112264 were identified as benign samples being predicted as “cancer”. GSE113486 dataset has no benign samples and there are only 3 wrong predictions. GSE113740 dataset has 450 wrong predictions, out of which

343 samples were benign in nature and 101 samples were liver cirrhosis, chronic hepatitis patients. 340 samples predictions were wrong in the dataset GSE122497, out of which 323 samples were of benign nature. 215 of 216 wrongly predicted samples of GSE137140 dataset were benign class. 157 out of 166, 487 out of 488 samples wrong predictions, in GSE164174 and GSE164174 datasets respectively were from benign samples. GSE73002 predicted 79 breast cancer samples as "non_cancer". In training dataset also, 625 out of 643 wrongly predicted samples were benign samples (**Fig. 5H**). Overall, the benign samples were segregated to "cancer" rather than "non_cancer" samples by the pan-cancer classifier (**Fig. 3E, Fig. 4C, Supp. Fig. 5J**)**.** Healthy control samples or samples from other diseases were predicted correctly as "non_cancer" class, as observed in the validation datasets.

Classifier performance increases when the benign samples were assigned in "cancer" class

Since benign samples affect the performance of the pan-cancer classifier, the distribution of "cancer", "non_cancer" and "benign" classes in the test datasets were observed through the t-SNE plots (**Supp. Fig. 6 A-J)**. It was observed that "benign" samples indeed form a distinct distribution that are close to cancer samples in many datasets. The classifier performance over test datasets were evaluated after assigning the "cancer" class to "cancer/benign" class that comprises of cancer and benign samples. The "non_cancer" class was restructured to "non_cancer/non_benign" samples that lacks the benign samples. The individual rules performances were visualized through the confusion matrices and it is observed that the performance of all the four rules improved when "benign" is included in the "cancer" class (**Supp. Fig.7**). The pan-cancer classifier performance increased in identifying the

“cancer/benign” class in comparison with the performance when benign is included in the “non_cancer” class (**Fig. 6A, 6B)** (**Supp. Table 5**). The total number of samples in the training datasets and the prediction results when the benign is included in “non_cancer” class (**Fig. 6C**) and when benign is included in “cancer/benign” class (**Fig. 6D**) suggest the training data classification also improved. The total number of test datasets classification results when benign samples are included in “non_cancer” class (**Fig. 6E**) and when benign samples are included in “cancer/benign” class (**Fig. 6F**) (**Supp. Table 6**) suggest overall improvement in classification results of pan-cancer classifier. The performance metrices improved for individual rules in test datasets when benign was considered in “cancer/benign” class, in comparison with the “non_cancer” class (**Supp. Fig. 8**).

The performance metrices of the test datasets when benign was categorized as cancer/benign plotted through the heatmap suggest the best performance of the pan-cancer classifier (**Fig. 6G**). The overall performance increased very significantly and the ROC plots denotes the excellent diagnostic potential of the classifier at single-sample level (**Fig. 6H**).

Taken together, the rules-based single-sample pan-cancer serum classifier developed in this study achieved higher performance over a massive sample size of 46349 samples from 18 independent studies. Although many transcriptome-based classifier models were identified in several studies, no study has identified pan-cancer serum-based classifier that were evaluated in such a larger sample size. The uniqueness of the single-sample pan-cancer classifier model identified in this study is that the classifier can predict the presence of cancer as well as pre-cancerous benign lesions containing neoplastic cells sample-wise and hence can be intervened

for screening programmes for further biochemical characterization of the suspected individuals with least invasive compared to the current available screening and diagnosis methods.

Discussion

Liquid biopsy allows a minimally invasive way of early diagnosis of cancer and thus can help in reduction of cancer-associated mortality. The existing methods of cancer detection involves the utilization of highly-invasive methods that are expensive and hence not suitable for large scale screening for cancer. The most important drawback of the existing liquid biopsy diagnosis methods that utilizes the cancer-associated mutations is the prevalence of false-positive results[27]. Cancer-associated databases such as TCGA has only the data derived from various cancer tissues and miRNA expression resource is needed for liquid biopsy in various cancer types. This study performed data mining and curation to build a web resource that have expression of miRNAs in 46349 clinical serum samples. This resource will be valuable for any other studies that need to estimate the expression of miRNAs in cancer and other disease conditions in the serum of clinical subjects.

Next, we proposed to identify the miRNA-based classifier that can accurately discriminate between “cancer” and “non_cancer” samples. There are limited or no studies that explored the possibility of a pan-cancer classifier that can identify the presence of cancer directly from serum. The existing studies utilizing machine learning approach to identify the miRNA-based classifier to discriminate between “cancer” and “non_cancer” classes have focused mostly on the correct prediction of cancer. It is equally important to correctly identify the other disease samples as “non_cancer” as the same miRNAs can be dysregulated in cancer as well as other pathological conditions. The classifier model evaluation is limited by the availability of suitable datasets with both the “cancer” and “non_cancer” class comprising other disease conditions. The conventional

ML classifier performances are affected by the class imbalance, normalization methods as the ML algorithms identify the “cancer” and “non_cancer” class-specific discriminatory features. To overcome the limitations of the conventional ML classifiers, this study utilized the approach of single-samples rules-based classification algorithms that relies on the relationship of miRNA pairs within the samples and predict the “cancer” or “non_cancer” class. The serum miRNA resource built in this study were utilized for the training of the classifier model and evaluation in test and validation datasets consisting of independent datasets comprising the samples belonging to different cancer types, non-cancer samples including several other diseases.

A total of eight miRNAs constituting 4 miRNA pairs were identified to be accurately discriminating between the “cancer” and “non_cancer” class. The voting-based approach was utilized for the prediction of the class in which more than 2 rules satisfying represents the “cancer” class, otherwise, “non_cancer” class. miR-4258 was observed to have very high expression only in serum but not in whole blood, denoting extra-blood origin in an individual study[28]. miR-4258 expression was observed to be upregulated in various cancers and play an oncogenic role by contributing to metastasis in liver cancer[29,30]. miR-4730 was observed to play a tumor suppressive role in liver cancer [31] and the lesser expression is associated with poorer survival in pancreatic cancer [32]. In our study, we identified the expression of oncogenic miR-4258 was observed to be higher than the expression of tumor-suppressive miR-4730 in “cancer” samples whereas the reverse trend was observed in “non_cancer” samples. It is speculated that the pathology associated with the cancer might lead to the induction of miR-4730 and reduction of miR-4730 and this change can reflect the tumor status. The second rule was miR663a > miR-1228-5p to be predicted as cancer. miR-663a was found to play a tumor promoting role in

nasopharyngeal carcinoma and its upregulation was observed to be associated with poorer prognosis. Interestingly, miR-663a is observed to play a tissue specific role in cancer as it promotes malignancy in prostate cancer while it suppresses in pancreatic cancer. It plays a major role related to inflammation genes such as TGFB1 [33]. miR-1228-5p is said to be enriched in exosomes and plays an important role in cell migration and proliferation by interacting with an enzyme DUSP22, which regulates cell proliferation. Down-regulation of DUSP22 was observed in the cells where miR-1228-5p was high in small cell lung cancer in-vivo studies, leading to increase in migration and proliferation of cancerous cells [34]. The role of miR-6787-5p, miR-6800-5p, miR-1469 and miR-1268a in cancer diagnosis is yet to be explored.

Initially the pan-cancer classifier performance was observed to be affected while predicting certain “non_cancer” samples that were misclassified as “cancer” samples. After exploring the distribution of incorrect predictions, it was observed those samples were majorly of benign nature. This is interesting in the perspective of early diagnosis applications as the screening of pre-cancerous/ benign samples that are associated with initiation of neoplastic growth and may have potential to turn to cancerous at the early stage is crucial for the improved diagnosis before the advent of the tumor proliferation. This classifier can be used as screening of the high-risk subjects who can be recommended for further clinical diagnosis cancer methods.

This study also has several limitations. Since the study is based on the availability of miRNA expression datasets from the clinical samples available in the public databases, it is restricted only to serum due to the increased availability. There are few reports that suggested the miRNA expression changes between serum, plasma and whole blood and hence this classifier cannot be

generalized for all the liquid biopsy miRNA expression. The availability of small-RNA sequencing datasets derived from the serum samples were limited and also the expression of the miRNAs was dependent on the depth of sequencing and library preparation. Hence, the future work is planned to design the custom panel for these miRNAs and identify the rules for the prediction of “cancer” and “non_cancer” samples from the serum samples.

Methods

Clinical datasets screening, selection and acquisition:

The public transcriptome databases such as Gene Expression Omnibus (GEO) and array express were queried with the keyword “serum miRNA” to retrieve the miRNA transcriptome datasets derived from the serum of the clinical subjects. This search resulted in 68,022 samples across 1495 datasets. The exclusion criteria include the samples that are derived from the cell-line models, non-serum tissue origin including those from exosomes and single-cell RNA-seq datasets. Only the datasets with minimal sample size of 500 was selected for train and test datasets. Duplicates or samples with insufficient metadata information were removed from the analysis. A total of 46,349 samples across 18 independent studies were further selected for the study (**Supp. Fig. 1**) (**Supp. Table 1**).

Study design:

The dataset GSE211692 was utilized as training dataset due to its diverse representation of 13 different cancer types and an adequate sample size for each cancer type, along with the non-cancer controls. The test datasets were independent datasets that comprises both the “cancer” and “non_cancer” classes within each dataset to evaluate the performance of the trained model. The validation datasets comprise of the samples that are having either “cancer” or “non_cancer” class of different disease types. The performance of the classifier in the identification of other disease classes as “non_cancer” class can be explored through the validation datasets.

Training dataset: GSE211692 (n = 16,190)

Test datasets: GSE122497, GSE106817, GSE137140, GSE113740, GSE112264, GSE139031, GSE113486, GSE164714, and GSE73002 (n = 25,508)

Validation datasets: GSE110317, GSE110651, GSE134108, GSE85589, GSE117064, GSE120584, GSE150693, and GSE140249 (n = 5,077)

The datasets included a variety of cancer types including, benign conditions, and non-cancer diseases such as Autoimmune Hepatitis, Alzheimer's disease (AD), Vascular Dementia (VD), and Mild Cognitive Impairment (MCI). (**Supp. Table 1**).

Dataset retrieval and pre-processing:

The miRNA expression matrices and the metadata information about the clinical subjects were retrieved through the R package GEOquery[35]. In training data, only the miRNAs that are expressed in at least 10% of samples were retained. This is done to ensure that the expression of miRNAs should be at higher level in serum and hence after the training of the classifier model, the expression the miRNAs must be there in test and validation datasets too. A total of 211 miRNAs with higher expression was remained for the further analysis. Batch effect correction was not required in the case of test datasets as the classifier was evaluated as single sample level rather than the dataset level.

Training of the pan-cancer classifier:

The rules-based single-sample classifier was developed based on the pairwise comparisons between the miRNAs. The top scoring pairs (TSPs) for the samples belonging to the "cancer" and "non_cancer" class were identified. If there are "N" number of miRNAs in the expression

matrix, the total number of possible pairs per class for the first miRNA will be N-1. The total possible number of pairs per class will be N(N-1)/2 pairs. For “m” number of samples, the total number of possible pairs per class will be m*N(N-1)/2 pairs. There are a total 16190 samples in the training dataset and the number of highly expressed miRNAs after the filtration step was 211 miRNAs. So, for each class there will be 211*210/2 = 22155 pairs was calculated. For “cancer” and “non_cancer” the combined restricted pairs will be 22155+22155 = 44310 pairs. For 16190 samples, a total of 3.5*10^8 pairs were calculated for “cancer” class and 3.5*10^8 pairs were calculated for “non_cancer” class leading to the analysis of around 7.17*10^8 pairs. The top scoring pairs for each class was calculated by the R packages, switchBox[36] and multiclassPairs[37]. A total of 4 rules were identified based on the scores that gives the idea about the trajectory specific for the particular class. The classification rules were determined by the miRNA expression relationships:

For example, if the expression of miRNA 4258 > miRNA 4730, then the sample will be classified as “cancer”. If the expression of miRNA 4730 > miRNA 4258, the sample will be classified as “non_cancer”. The miRNA names are mentioned only with the number and 5p or 3p if applicable for the sake of better annotation during the training.

A total of 4 TSPs were identified and a voting-based approach was implemented to identify the class of the sample with higher accuracy. If more than two rules out of 4 rules for “cancer” class were satisfied, then it will be classified as “cancer”. If not, it will be classified as “non_cancer”.

Performance Evaluation:

The Evaluation of the performance of the pan-cancer serum classifier was performed by calculating the performance metrices such as Sensitivity, specificity, accuracy, Precision and F1-score. The confusion matrices were plotted to depict the relationship between the correct predictions of “cancer” class - True-positive (TP) – “cancer” identified as “cancer”, True-negative (TN) – “non_cancer” predicted as “non_cancer” and the incorrect predictions – False-postive (FP) – “non_cancer” predicted as “cancer”, False-negative (FN) – “cancer” predicted as “non_cancer”.

Sensitivity measures the classifier ability to predict the “cancer” samples correctly. Higher sensitivity denotes the lesser “cancer” samples are misclassified as “non_cancer”.

Sensitivity = TP/(TP+FN)

Specificity measures the ability of the classifier to identify the non-cancer samples correctly as “non_cancer”. A high specificity denotes lesser “non_cancer” samples misclassified as “cancer”.

Specificity = TN/(TN+FP)

Accuracy measures the proportion of samples predicted correctly out of all the samples tested. The higher value of accuracy represents the overall correctness of the classifier.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision measures how many of the predicted cancer cases were actually cancer. Higher precision means fewer “non_cancer” samples were misclassified as cancer.

Precision = TP/(TP+FP)

Recall measures the ability of the classifier to correctly identify the positive instances. Higher recall value depicts the lesser false negatives.

Recall = TP/(TP+FN)

F1-score measures the balance between the sensitivity and precision. It is important when the dataset is affected by class imbalance.

F1-score = 2 * [(Precision * Sensitivity)/ (Precision + Sensitivity)]

The confusion matrices were plotted in form of heatmap representing the total number of samples and all the metrices were calculated and plotted using the Machine learning R package caret. For the better visualization all the metrices are given as percentages in the plot.

ROC analysis:

Receiver Operating characteristics (ROC) curve plots were plotted to represent the relation between the true positive rate (y-axis) and the false positive rate (x-axis). ROC analysis was performed and visualized through the R package pROC[38].

The classifier's robustness was assessed across independent test and validation datasets. The impact of benign samples on classification performance was analyzed by reassigning the "cancer" class to "cancer/benign" class by including the benign samples and "non_cancer" class to "non_cancer/non_benign" by excluding benign samples.

Visualization methods:

For the distribution of samples in each dataset, t-SNE plots were plotted through the R package "Rtsne". For better understanding all the "cancer" class was represented in red color and "non_cancer" class was represented in blue color. The distribution plots for the training, test and validation datasets were plotted as lollipop and tile plots that tells the relationship between the number of samples in each disease state. The expression of eight miRNAs forming the four rules were visualized using beeswarm plots plotted through the R package "ggbeeswarm". Bee-swarm plots used for the better visualization in distinguishing individual data points and prevent over-plotting, especially in the datasets with larger sample sizes. The incorrect predictions across the training, test and validation datasets were plotted as bar plots denoting the number of correct and wrong predictions in each dataset. The R packages ggplot2, dplyr, tidyverse were used for the plots and data-frame operations.

All the analysis was performed using R 4.4.1 version in a computational workstation with I9 processor with 32 threads and 128GB RAM physical memory.

Data and code availability:

All the datasets used in the study are publicly available from the GEO database. The codes used for the analyses performed in the study can be requested upon the contacting the corresponding author.

References:

1. Bray, F., *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **74**, 229-263 (2024).
2. Bray, F., Laversanne, M., Weiderpass, E. & Soerjomataram, I. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* **127**, 3029-3030 (2021).
3. Siegel, R.L., Miller, K.D., Wagle, N.S. & Jemal, A. Cancer statistics, 2023. *CA Cancer J Clin* **73**, 17-48 (2023).
4. Hashim, D., *et al.* The global decrease in cancer mortality: trends and disparities. *Ann Oncol* **27**, 926-933 (2016).
5. Ma, L., *et al.* Liquid biopsy in cancer current: status, challenges and future prospects. *Signal Transduct Target Ther* **9**, 336 (2024).
6. Weber, J.A., *et al.* The microRNA spectrum in 12 body fluids. *Clin Chem* **56**, 1733-1741 (2010).
7. Brase, J.C., Wuttig, D., Kuner, R. & Sultmann, H. Serum microRNAs as non-invasive biomarkers for cancer. *Mol Cancer* **9**, 306 (2010).
8. Matsuzaki, J., *et al.* Prediction of tissue-of-origin of early stage cancers using serum miRNomes. *JNCI Cancer Spectr* **7**(2023).
9. Sudo, K., *et al.* Development and Validation of an Esophageal Squamous Cell Carcinoma Detection Model by Large-Scale MicroRNA Profiling. *JAMA Netw Open* **2**, e194573 (2019).
10. Suzuki, K., *et al.* Circulating serum miRNAs predict response to platinum chemotherapy in high-grade serous ovarian cancer. *Cancer Med* **13**, e70251 (2024).
11. Yokoi, A., *et al.* Integrated extracellular microRNA profiling for ovarian cancer screening. *Nat Commun* **9**, 4319 (2018).
12. Asakura, K., *et al.* A miRNA-based diagnostic model predicts resectable lung cancer in humans with high accuracy. *Commun Biol* **3**, 134 (2020).
13. Yamamoto, Y., *et al.* Highly Sensitive Circulating MicroRNA Panel for Accurate Detection of Hepatocellular Carcinoma in Patients With Liver Disease. *Hepatol Commun* **4**, 284-297 (2020).
14. Urabe, F., *et al.* Large-scale Circulating microRNA Profiling for the Liquid Biopsy of Prostate Cancer. *Clin Cancer Res* **25**, 3016-3025 (2019).
15. Ohno, M., *et al.* Assessment of the Diagnostic Utility of Serum MicroRNA Classification in Patients With Diffuse Glioma. *JAMA Netw Open* **2**, e1916953 (2019).
16. Usuba, W., *et al.* Circulating miRNA panels for specific and early detection in bladder cancer. *Cancer Sci* **110**, 408-419 (2019).
17. Abe, S., *et al.* A novel combination of serum microRNAs for the detection of early gastric cancer. *Gastric Cancer* **24**, 835-843 (2021).
18. Shimomura, A., *et al.* Novel combination of serum microRNA for detecting breast cancer in the early stage. *Cancer Sci* **107**, 326-334 (2016).
19. Shiino, S., *et al.* Serum miRNA-based Prediction of Axillary Lymph Node Metastasis in Breast Cancer. *Clin Cancer Res* **25**, 1817-1827 (2019).
20. Satomi-Tsushita, N., *et al.* Serum microRNA-based prediction of responsiveness to eribulin in metastatic breast cancer. *PLoS One* **14**, e0222024 (2019).

21. Sato, J., *et al.* Brain metastasis-related microRNAs in patients with advanced breast cancer. *PLoS One* **14**, e0221538 (2019).
22. Lee, J., *et al.* Identification of Circulating Serum miRNAs as Novel Biomarkers in Pancreatic Cancer Using a Penalized Algorithm. *Int J Mol Sci* **22**(2021).
23. Sonoda, T., *et al.* Serum MicroRNA-Based Risk Prediction for Stroke. *Stroke* **50**, 1510-1518 (2019).
24. Shigemizu, D., *et al.* Risk prediction models for dementia constructed by supervised principal component analysis using miRNA expression data. *Commun Biol* **2**, 77 (2019).
25. Shigemizu, D., *et al.* Prognosis prediction model for conversion from mild cognitive impairment to Alzheimer's disease created by integrative analysis of multi-omics data. *Alzheimers Res Ther* **12**, 145 (2020).
26. Saeki, C., *et al.* Identification of circulating microRNAs as potential biomarkers for hepatic necroinflammation in patients with autoimmune hepatitis. *BMJ Open Gastroenterol* **9**(2022).
27. Batool, S.M., *et al.* The Liquid Biopsy Consortium: Challenges and opportunities for early cancer detection and monitoring. *Cell Rep Med* **4**, 101198 (2023).
28. Pascut, D., *et al.* A comparative characterization of the circulating miRNome in whole blood and serum of HCC patients. *Sci Rep* **9**, 8265 (2019).
29. Yokota, Y., *et al.* Serum exosomal miR-638 is a prognostic marker of HCC via downregulation of VE-cadherin and ZO-1 of endothelial cells. *Cancer Sci* **112**, 1275-1288 (2021).
30. Zhong, S., *et al.* MicroRNA expression profiles of drug-resistance breast cancer cells and their exosomes. *Oncotarget* **7**, 19601-19609 (2016).
31. Furuke, H., *et al.* miR□4730 suppresses the progression of liver cancer by targeting the high mobility group A1 pathway. *Int J Mol Med* **49**(2022).
32. Yang, W., Ju, H.Y. & Tian, X.F. Hsa-miR-4730 as a new and potential diagnostic and prognostic indicators for pancreatic cancer. *Eur Rev Med Pharmacol Sci* **24**, 8801-8811 (2020).
33. Tian, W., *et al.* miR663a□TTC22V1 axis inhibits colon cancer metastasis. *Oncol Rep* **41**, 1718-1728 (2019).
34. Mu, X., *et al.* Exosomal miR-1228-5p down-regulates DUSP22 to promotes cell proliferation and migration in small cell lung cancer. *Life Sci* **351**, 122787 (2024).
35. Davis, S. & Meltzer, P.S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846-1847 (2007).
36. Afsari, B., Fertig, E.J., Geman, D. & Marchionni, L. switchBox: an R package for k-Top Scoring Pairs classifier development. *Bioinformatics* **31**, 273-274 (2015).
37. Marzouka, N.A. & Eriksson, P. multiclassPairs: an R package to train multiclass pair-based classifier. *Bioinformatics* **37**, 3043-3044 (2021).
38. Robin, X., *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

Acknowledgement:

This work was supported by the intramural research grant from IISER Bhopal and the support for the computational workstation from Indian Institute of Science Education and Research, Bhopal (IISER Bhopal). We thank all the authors who have deposited the miRNA transcriptome datasets in public databases.

Author information:

Authors and Affiliations:

Indian Institute of Science Education and Research, Bhopal (IISER Bhopal)

Pandikannan Krishnamoorthy

Madhavan Parthasarathy

Nilanjana Das

Himanshu Kumar

All India Institute of Medical Sciences, Bhopal (AIIMS Bhopal)

Ashok Kumar

Vikas Gupta

Saikat Das

Frontiers Research Center, Osaka University

Himanshu Kumar

Contributions:

P.K and H.K. conceived the project. P.K, M.P, N.D and ASR contributed in the datasets screening. P.K. optimized and formulated the methodology and wrote the codes for the analysis. P.K and M.P. did the formal analysis and visualization. M.P and P.K annotated and developed the web resource. P.K, M.P, N.D and H.K wrote the original draft of the paper. A.S, V.G and S.D contributed to the review of the study design, P.K, M.P, N.D, ASR and H.K reviewed and edited the paper. H.K. supervised the entire project. All authors read and approved the final paper.

Ethics declarations

All the analyses are made from the publicly available clinical transcriptome datasets. No patients were recruited nor any experiment is performed on clinical subjects and hence requirement of Institute Ethical committee is not required for this study.

H.K and P.K has a patent registered related to this work.

Figure Legends

Fig. 1:

Schematic Representation of Selection of Serum miRNA datasets, samples distribution and Rules-Based single sample Classifier Model Construction.

(A) Overview of serum samples distribution of "cancer" and "non_cancer." The "cancer" samples consist of 13 cancer types and the "non_cancer" samples consist of healthy and other disease as depicted in A. The number in parentheses represent number of samples. Overall 46,349 samples were included that were split into, Training data (16,190 samples) to build the rules-based machine learning classifier model and the resulting model is tested in, Test datasets (25,082 samples) obtained from 9 independent datasets contain both cancer and non_cancer classs. The robustness of the model as single sample classifier in identifying even cancer and non_cancer samples separately is validated in (iii) 8 independent Validation datasets (5077 samples) that comprise either "cancer" or "non_cancer" that includes autoimmune diseases, neurological disorders and cerebrovascular disorders. **(B)** Superiority of the single-sample classifier (SSC) approach over the conventional machine learning (ML) classifier (CMC) is depicted in three cases of equal sample size between the classs (Case 1), class imbalance (Case 2) and the need of equal number of classes (Case 3).

Fig. 2:

Identification of top miRNAs rules combination to classify "cancer" and "non_cancer" serum samples

(A) Lollipop plot illustrates the distribution of different types of "cancer" (red) and "non_cancer"

(blue) types present in the training data GSE211692. (B) Rules-based training approach is illustrated in the schematic. miRNA expression processing phase comprises filtering, generation of top-scoring pairs (TSPs), training process of classifiers using one versus rest scheme and calculation of scores of each TSP pair. The top k rules for each class were identified. (C) The segregation of the “cancer” (red) and “non_cancer” (blue) classes are illustrated through the T-distributed Stochastic Neighbor Embedding (t-SNE) plot. (D) The Final four score rules for each class identified were denoted and their comparison rules in “cancer” and “non_cancer” classes are denoted. (E) The confusion matrices denote the relation between the predicted and the actual classes for rules: (E) 4258 > 4730, (F) 663a > 1228_5p, (G) 6787_5p > 4800_5p and (H) 1469 > 1268a.

<u>Fig. 3:</u>

Construction of pan-cancer serum miRNA rules-based classifier

(A) The expression of eight miRNAs that form the four rules across the “cancer” and “non_cancer” classes were plotted through the bee-swarm plots where red denotes “cancer” and blue denotes “non_cancer” class. (B) The individual rules and their classification metrics such as Specificity, Sensitivity, Recall, Precision, F1 value and accuracy were plotted as heatmap. (C) Schematic of the classifier construction incorporating the voting-based classification in which the samples are predicted as cancer if more than two rules are satisfied and “non_cancer” if less than 2 or equal to 2 satisfied rules. (D) Confusion matrix denotes the relationship between the predicted and true classes when all four rules are combined as a classifier. (E) Heatmap denotes the classification performance of all the four rules that are selected as classifier. Prediction Scores for each sample were plotted in red and green for “cancer” and “non_cancer”

respectively. The distribution of the disease types and the corresponding predictions were depicted at the bottom of the heatmap.

<u>Fig. 4:</u>

Evaluation of pan-cancer classifier performance in independent test datasets

(A) Tiles plot denotes the distribution and all the types of samples present across all the 9 test datasets and the tile color represents the class- red for “cancer” and blue for “non_cancer”. (B) The performance of the classifier in 9 individual datasets were plotted as confusion matrices that denote the relationship between the predictions (x-axis) and the true label (y-axis). (C) The heatmap plot depicts the prediction results of all the 25082 samples across 9 independent test datasets comprising different disease states and prediction Scores for each sample were plotted in orange and light green for “cancer” and “non_cancer” respectively. (D) The classifier performance in individual datasets and their classification metrices such as Specificity, Sensitivity, Recall, Precision, F1 value and accuracy were plotted as heatmap. (E) The Receiver operator curves (ROC) to evaluate the diagnostic potential by comparing the true positive rate and false positive rate for all the dataset and the area under curve values for each of the datasets were illustrated.

<u>Fig. 5:</u>

Pan-cancer classifier robustness as single-sample classifier and its influence on benign samples:

(A) Tiles plot denotes the distribution and the number of all types of samples present across all the 8 validation datasets and the tile color represents the class-red for “cancer” and blue for “non_cancer.” (B) Evaluation of pan-cancer classifier performance in individual datasets were

plotted as confusion matrices that denote the relationship between the predicted and true labels (C) The Bar plot denotes the number of samples predicted correct and incorrect by the classifier in the validation datasets where the light green represents the correct predictions and orange represents the incorrect prediction. The values on each bar represent the total number of samples in each dataset. (D) The tile plot depicts the distribution of sample types that are predicted incorrectly in the validation datasets. All the benign samples in the GSE110317 are predicted as cancer. (E) The Bar plot denotes the number of samples predicted correct and incorrect by the classifier in the test datasets where the light green represents the correct predictions and orange represents the incorrect prediction. The values on each bar represent the total number of samples in each dataset. (F) The tile plot depicts the distribution of sample types that are predicted incorrectly in the test datasets. (G) The Bar plot denotes the number of samples predicted correct and incorrect by the classifier in the training data where the light green represents the correct predictions and orange represents the incorrect prediction. The values on each bar represent the total number of samples in each dataset. (H) The bar plot depicts the distribution of sample types that are predicted incorrectly in the training dataset.

<u>Fig. 6:</u>

Pan-cancer classifier predicts the cancer and pre-cancerous benign conditions with highest accuracy:

The confusion matrices of individual test datasets where (A) benign cases are considered “non_cancer” and (B) “cancer/benign” classes were depicted. The confusion matrices of training data when (C) benign cases are considered “non_cancer” and (D) “cancer/benign” classes, test datasets (all samples from each datasets) as (E) “non_cancer” and (F) “cancer/benign” depicts

the improved classification of “cancer/benign” class. (G) The classifier performance metrics when benign cases are considered “cancer/benign” class is depicted as heatmap. (H) The ROC plots depict the highly accurate diagnostic performance in differentiating “cancer/benign” and “cancer/non-benign” classes.

Supplementary Figures Legend:

Supp. Fig.1:

Serum samples and selection criteria: Serum miRNA datasets sample-wise selection strategy and criteria following PRISMA guidelines.

Supp. Fig.2:

Sample distribution in test datasets: t-distributed Stochastic Neighbour Embedding (t-SNE) plots depicting the distribution of samples in test datasets between cancer and non_cancer classs in (A) GSE106817, (B) GSE112264, (C) GSE113486, (D) GSE113740, (E) GSE122497, (F) GSE137140, (G) GSE139031, (H) GSE164174, (I) GSE73002.

Supp. Fig. 3:

Expression of classifier miRNAs in test datasets: Expression of 8 miRNAs constituting the classifier rules in test datasets (A) GSE106817, (B) GSE112264, (C) GSE113486, (D) GSE113740, (E) GSE122497, (F) GSE137140, (G) GSE139031, (H) GSE164174, (I) GSE73002. Red denotes the cancer class while the blue denotes the non_cancer class. y-axis denotes the normalized expression values and x-axis denotes the class names.

Supp. Fig. 4:

Classification of true and predicted cancer and non_cancer in test datasets: Confusion matrices depicting the comparison between the classifier results and the actual labels as explained in (A), for the 4 rules individually in all the 9 test datasets (B-J).

Supp. Fig. 5:

Classification of true and predicted cancer and non_cancer in test datasets: Confusion matrices depicting the comparison between the classifier results and the actual labels as explained in (A), for the 4 rules individually in all the 8 Validation datasets (B-I). (J) The heatmap depicts the prediction results of all the samples that are part of validation datasets comprising different disease types.

Supp. Fig. 6:

Benign Sample distribution in test datasets: t-distributed Stochastic Neighbour Embedding (t-SNE) plots depicting the distribution of samples in test datasets between "cancer/benign" and "non_cancer" classes in (A) GSE106817, (B) GSE112264, (C) GSE113486, (D) GSE113740, (E) GSE122497, (F) GSE137140, (G) GSE139031, (H) GSE164174, (I) GSE73002. The benign samples form a distinct distribution.

Supp. Fig. 7:

Classification of true and predicted cancer, benign, and non_cancer in test datasets: Confusion matrices depicting the comparison between the classifier results and the actual labels between cancer/benign as one class and non_cancer/non_benign as another class as explained in (A), for the 4 rules individually in all the 9 test datasets (B-J).

Supp. Fig. 8:

Model Performance after considering benign and pre-cancerous lesion as cancer: The classifier performance metrics such as Specificity, Sensitivity, Recall, Precision, F1 and Accuracy are plotted as heatmap. The comparison between the performance when benign is considered as non_cancer (A) and when benign is considered as cancer (B) is depicted.

Supplementary Tables:

Supp. Table 1:

The training, test and validation datasets used in the study.

Supp. Table 2:

Prediction results of training dataset by the serum pan-cancer classifier.

Supp. Table 3:

Prediction results of test datasets by the serum pan-cancer classifier.

Supp. Table 4:

Prediction results of validation datasets by the serum pan-cancer classifier.

Supp. Table 5:

Prediction results of training dataset when benign is categorized in cancer/benign class by the pan-cancer classifier.

Supp. Table 6:

Prediction results of test datasets when benign is categorized in cancer/benign class by the pan-cancer classifier.

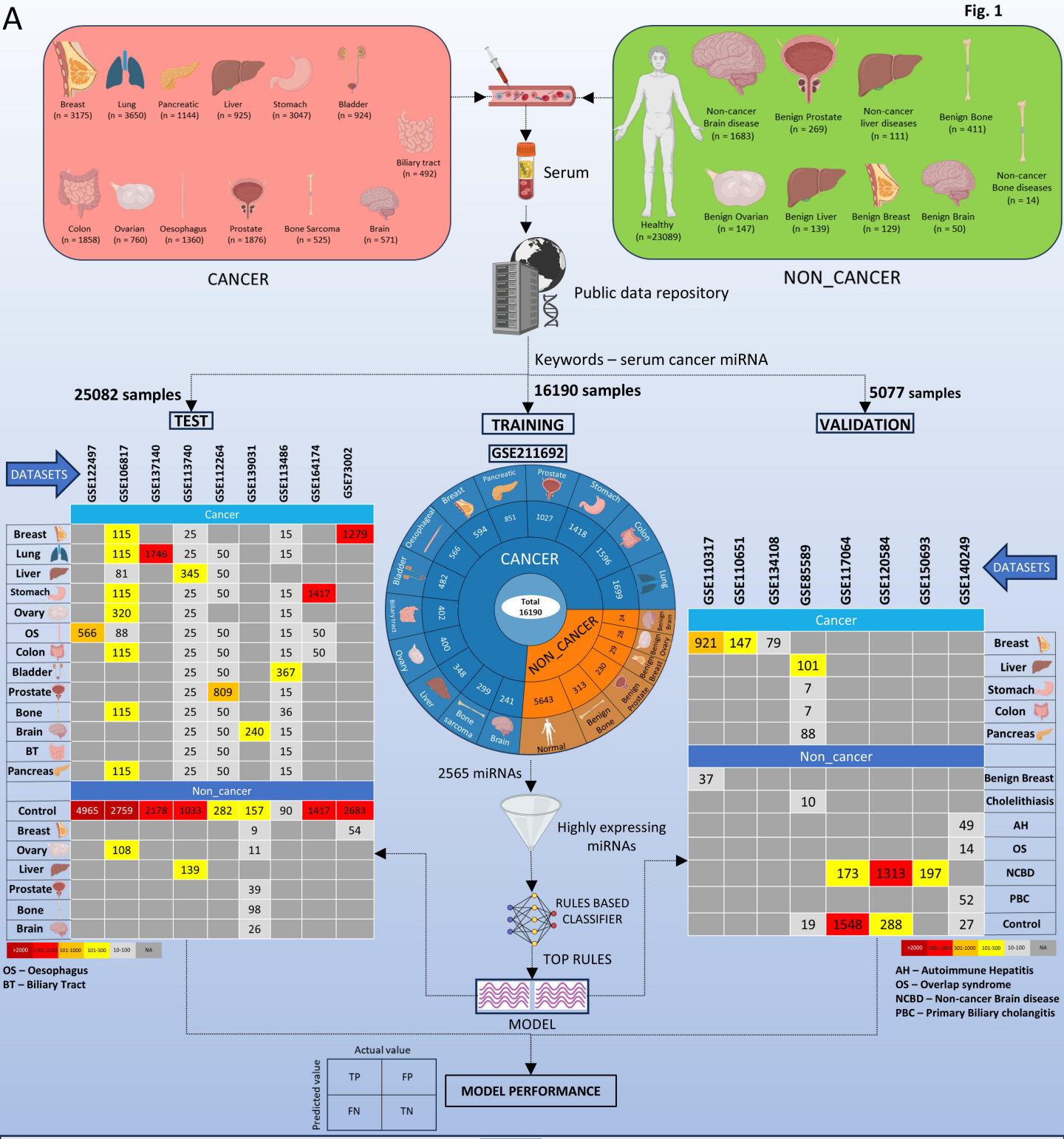

A
Fig. 1
Breast (n = 3175)
Lung (n = 3650)
Pancreatic (n = 1144)
Liver (n = 925)
Stomach (n = 3047)
Bladder (n = 924)
Biliary tract (n = 492)
Colon (n = 1858)
Ovarian (n = 760)
Oesophagus (n = 1360)
Prostate (n = 1876)
Bone Sarcoma (n = 525)
Brain (n = 571)
CANCER
Serum
Healthy (n =23089)
Non-cancer Brain disease (n = 1683)
Benign Prostate (n = 269)
Non-cancer liver diseases (n = 111)
Benign Bone (n = 411)
Benign Ovarian (n = 147)
Benign Liver (n = 139)
Benign Breast (n = 129)
Benign Brain (n = 50)
Non-cancer Bone diseases (n = 14)
NON_CANCER
Public data repository
Keywords – serum cancer miRNA
25082 samples
TEST
16190 samples
TRAINING
5077 samples
VALIDATION
DATASETS
GSE122497
GSE106817
GSE137140
GSE113740
GSE112264
GSE139031
GSE113486
GSE164174
GSE73002
Cancer
Breast 115 25 15 1279
Lung 115 1746 25 50 15
Liver 81 345 50
Stomach 115 25 50 15 1417
Ovary 320 25 15
OS 566 88 25 50 15 50
Colon 115 25 50 15 50
Bladder 25 50 367
Prostate 25 809 15
Bone 115 25 50 36
Brain 25 50 240 15
BT 25 50 15
Pancreas 115 25 50 15
Non_cancer
Control 4965 2759 2178 1033 282 157 90 1417 2683
Breast 9 54
Ovary 108 11
Liver 139
Prostate 39
Bone 98
Brain 26
>2000 1001-2000 501-1000 101-500 10-100 NA
OS – Oesophagus
BT – Biliary Tract
GSE211692
CANCER
Total 16190
NON_CANCER
Breast 594
Pancreatic 851
Prostate 1027
Stomach 1418
Colon 1596
Lung 1699
Oesophagal 566
Bladder 482
Biliary tract 202
Ovary 400
Liver 348
Bone sarcoma 299
Brain 241
Normal 5643
Benign Brain 24
Benign Ovary 28
Benign Breast 29
Benign Prostate 230
Benign Bone 313
2565 miRNAs
Highly expressing miRNAs
RULES BASED CLASSIFIER
TOP RULES
MODEL
GSE110317
GSE110651
GSE134108
GSE85589
GSE117064
GSE120584
GSE150693
GSE140249
DATASETS
Cancer
921 147 79
Breast
101
Liver
7
Stomach
7
Colon
88
Pancreas
Non_cancer
37
Benign Breast
10
Cholelithiasis
49
AH
14
OS
173 1313 197
NCBD
52
PBC
19 1548 288 27
Control
>2000 1001-2000 501-1000 101-500 10-100 NA
AH – Autoimmune Hepatitis
OS – Overlap syndrome
NCBD – Non-cancer Brain disease
PBC – Primary Biliary cholangitis
Actual value
Predicted value
TP
FP
FN
TN
MODEL PERFORMANCE
B
Conventional ML Classifier
miRNA_ID N1 N2 N3 N4 N5 C1 C2 C3 C4
miR1
miR2
miR3
miR4
miRn
Class Non_cancer
Class Cancer
ML algorithms
Discriminatory features between the classes
Training
miRNA_ID N1 C1 N2 N3 C2 C3 N4 C4 N5
miR1
miR2
miR3
miR4
miRn
Scales
0 1
Upregulated
Downregulated
CLASSIFIER
Performance evaluation
Test
Performance evaluation
Single sample Classifier
Sample 1
miR1
miR1 vs miR2
miR2
miR1 vs miRn
miR2 vs miRn
miRn
N-1
N-2
N-n = 0
= N X N-1/2
N = Total library pairs in a sample
m = Total number of samples
Total possible pairs per class for m samples = m(N X N-1/2)
Evaluate
Evaluate
Evaluate
Rule 1: miRNA1; miRNA x
Rule 2: miRNA2; miRNA y
Rule n: miRNA-X; miRNA-Y
TRUE
FALSE
TRUE
FALSE
TRUE
FALSE
Rules-based classifier votes
Case 1
miRNA_ID c11 c12 n11 c13 n12 c14 n13 n14
miR1
miR2
miR3
miR4
miRn
✓ CMC Actual performance of the classifier
✓ SSC Actual performance of the classifier
Case 2
miRNA_ID c11 c12 n11 c13 n12 c14 c15 c16 n13
miR1
miR2
miR3
miR4
miRn
✗ CMC Imbalance Datasets affect classifier performance
✓ SSC Imbalance Datasets will not affect classifier performance
Case 3
miRNA_ID c11 c12 c13 c14 c15
miR1
miR2
miR3
miR4
miRn
✗ CMC Test dataset should have both classes
✓ SSC Test dataset need not have both classes

Fig. 2

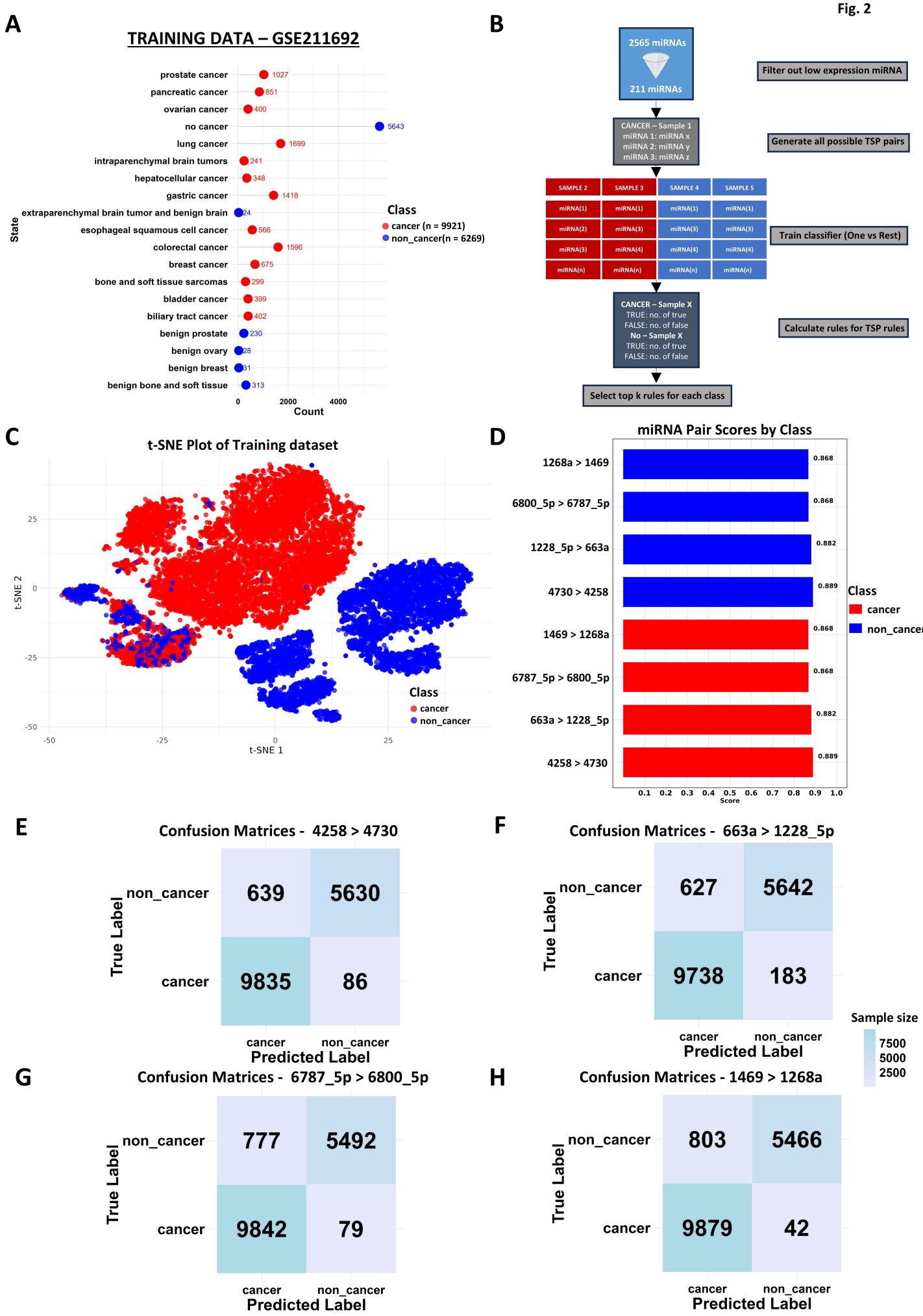

A
TRAINING DATA – GSE211692
prostate cancer 1027
pancreatic cancer 851
ovarian cancer 400
no cancer 5643
lung cancer 1699
intraparenchymal brain tumors 241
hepatocellular cancer 348
gastric cancer 1418
extraparenchymal brain tumor and benign brain 24
esophageal squamous cell cancer 566
colorectal cancer 1596
breast cancer 675
bone and soft tissue sarcomas 299
bladder cancer 399
biliary tract cancer 402
benign prostate 230
benign ovary 28
benign breast 31
benign bone and soft tissue 313
State
Count
Class
cancer (n = 9921)
non_cancer(n = 6269)
B
2565 miRNAs
211 miRNAs
Filter out low expression miRNA
CANCER – Sample 1
miRNA 1: miRNA x
miRNA 2: miRNA y
miRNA 3: miRNA z
Generate all possible TSP pairs
SAMPLE 2
SAMPLE 3
SAMPLE 4
SAMPLE 5
Train classifier (One vs Rest)
CANCER – Sample X
TRUE: no. of true
FALSE: no. of false
No – Sample X
TRUE: no. of true
FALSE: no. of false
Calculate rules for TSP rules
Select top k rules for each class
C
t-SNE Plot of Training dataset
t-SNE 1
t-SNE 2
Class
cancer
non_cancer
D
miRNA Pair Scores by Class
1268a > 1469 0.868
6800_5p > 6787_5p 0.868
1228_5p > 663a 0.882
4730 > 4258 0.889
1469 > 1268a 0.868
6787_5p > 6800_5p 0.868
663a > 1228_5p 0.882
4258 > 4730 0.889
Score
Class
cancer
non_cancer
E
Confusion Matrices - 4258 > 4730
non_cancer 639 5630
cancer 9835 86
True Label
Predicted Label
F
Confusion Matrices - 663a > 1228_5p
non_cancer 627 5642
cancer 9738 183
Sample size
7500
5000
2500
G
Confusion Matrices - 6787_5p > 6800_5p
non_cancer 777 5492
cancer 9842 79
H
Confusion Matrices - 1469 > 1268a
non_cancer 803 5466
cancer 9879 42

Fig. 3

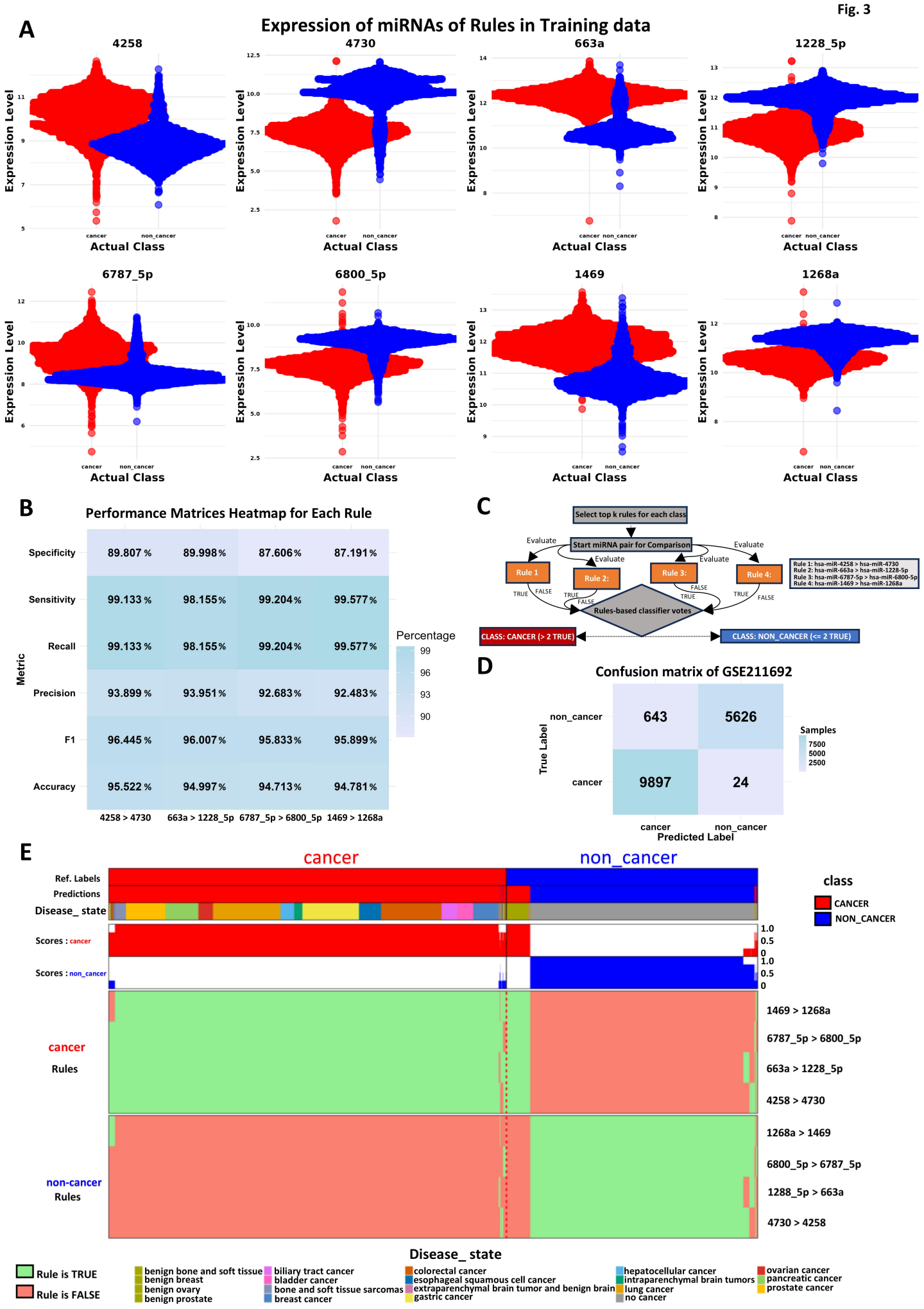

A
Expression of miRNAs of Rules in Training data
4258
4730
663a
1228_5p
6787_5p
6800_5p
1469
1268a
Expression Level
Actual Class
cancer
non_cancer
B
Performance Matrices Heatmap for Each Rule
Metric
Specificity
Sensitivity
Recall
Precision
F1
Accuracy
89.807 % 89.998 % 87.606 % 87.191 %
99.133 % 98.155 % 99.204 % 99.577 %
99.133 % 98.155 % 99.204 % 99.577 %
93.899 % 93.951 % 92.683 % 92.483 %
96.445 % 96.007 % 95.833 % 95.899 %
95.522 % 94.997 % 94.713 % 94.781 %
4258 > 4730
663a > 1228_5p
6787_5p > 6800_5p
1469 > 1268a
Percentage
99
96
93
90
C
Select top k rules for each class
Start miRNA pair for Comparison
Evaluate
Rule 1
Rule 2:
Rule 3:
Rule 4:
TRUE
FALSE
Rules-based classifier votes
CLASS: CANCER (> 2 TRUE)
CLASS: NON_CANCER (<= 2 TRUE)
Rule 1: hsa-miR-4258 > hsa-miR-4730
Rule 2: hsa-miR-663a > hsa-miR-1228-5p
Rule 3: hsa-miR-6787-5p > hsa-miR-6800-5p
Rule 4: hsa-miR-1469 > hsa-miR-1268a
D
Confusion matrix of GSE211692
True Label
non_cancer
cancer
643
5626
9897
24
Predicted Label
Samples
7500
5000
2500
E
cancer
non_cancer
Ref. Labels
Predictions
Disease_ state
Scores : cancer
Scores : non_cancer
class
CANCER
NON_CANCER
cancer Rules
non-cancer Rules
1469 > 1268a
6787_5p > 6800_5p
663a > 1228_5p
4258 > 4730
1268a > 1469
6800_5p > 6787_5p
1288_5p > 663a
4730 > 4258
Disease_ state
Rule is TRUE
Rule is FALSE
benign bone and soft tissue
benign breast
benign ovary
benign prostate
biliary tract cancer
bladder cancer
bone and soft tissue sarcomas
breast cancer
colorectal cancer
esophageal squamous cell cancer
extraparenchymal brain tumor and benign brain
gastric cancer
hepatocellular cancer
intraparenchymal brain tumors
lung cancer
no cancer
ovarian cancer
pancreatic cancer
prostate cancer

Fig. 4

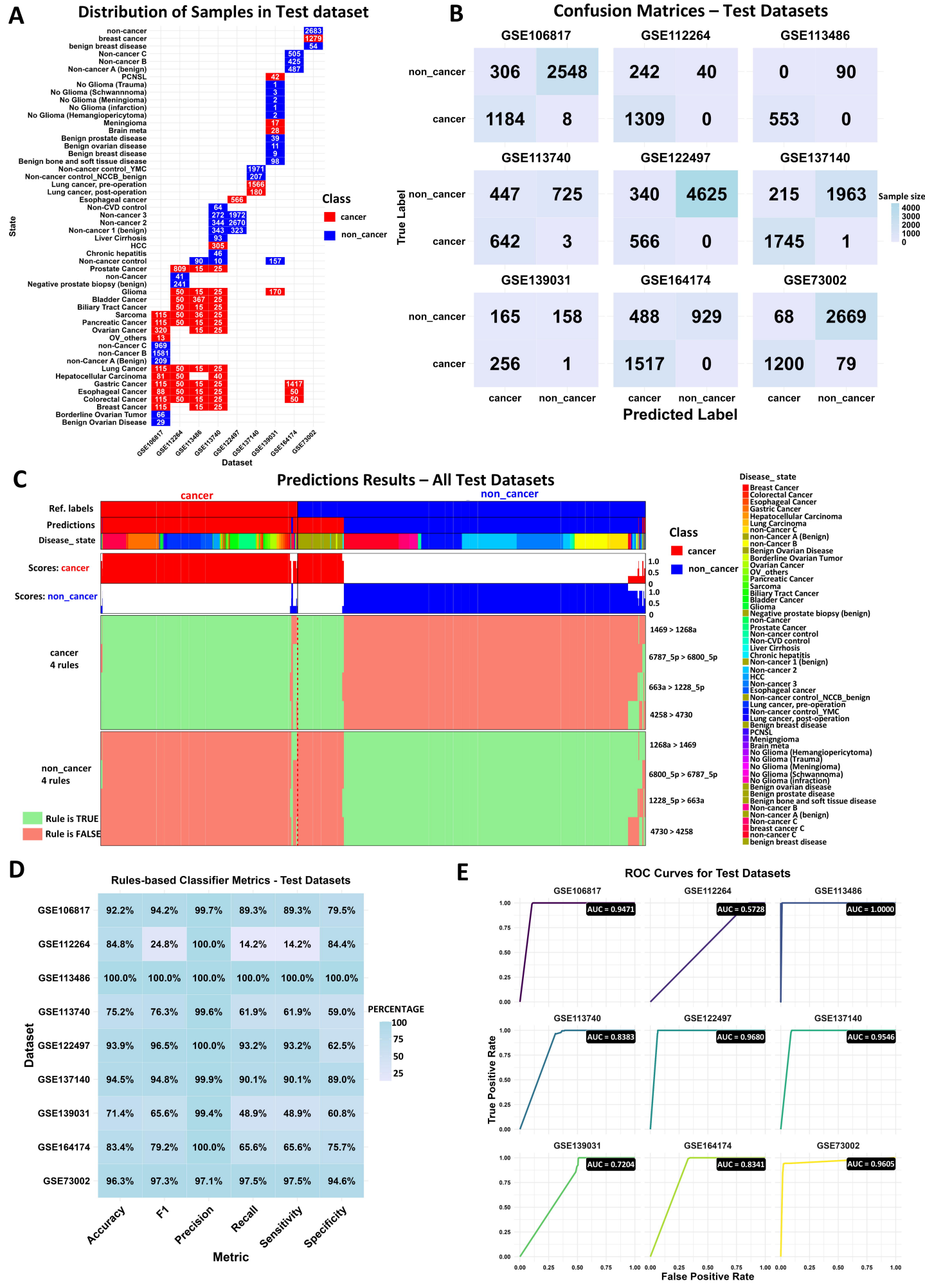

A
Distribution of Samples in Test dataset
State
Dataset
Class
cancer
non_cancer
B
Confusion Matrices – Test Datasets
GSE106817
non_cancer 306 2548
cancer 1184 8
GSE112264
non_cancer 242 40
cancer 1309 0
GSE113486
non_cancer 0 90
cancer 553 0
GSE113740
non_cancer 447 725
cancer 642 3
GSE122497
non_cancer 340 4625
cancer 566 0
GSE137140
non_cancer 215 1963
cancer 1745 1
GSE139031
non_cancer 165 158
cancer 256 1
GSE164174
non_cancer 488 929
cancer 1517 0
GSE73002
non_cancer 68 2669
cancer 1200 79
True Label
Predicted Label
Sample size
C
Predictions Results – All Test Datasets
cancer
non_cancer
Ref. labels
Predictions
Disease_ state
Scores: cancer
Scores: non_cancer
cancer 4 rules
non_cancer 4 rules
1469 > 1268a
6787_5p > 6800_5p
663a > 1228_5p
4258 > 4730
1268a > 1469
6800_5p > 6787_5p
1228_5p > 663a
4730 > 4258
Rule is TRUE
Rule is FALSE
Class
Disease_ state
D
Rules-based Classifier Metrics - Test Datasets
GSE106817 92.2% 94.2% 99.7% 89.3% 89.3% 79.5%
GSE112264 84.8% 24.8% 100.0% 14.2% 14.2% 84.4%
GSE113486 100.0% 100.0% 100.0% 100.0% 100.0% 100.0%
GSE113740 75.2% 76.3% 99.6% 61.9% 61.9% 59.0%
GSE122497 93.9% 96.5% 100.0% 93.2% 93.2% 62.5%
GSE137140 94.5% 94.8% 99.9% 90.1% 90.1% 89.0%
GSE139031 71.4% 65.6% 99.4% 48.9% 48.9% 60.8%
GSE164174 83.4% 79.2% 100.0% 65.6% 65.6% 75.7%
GSE73002 96.3% 97.3% 97.1% 97.5% 97.5% 94.6%
Accuracy
F1
Precision
Recall
Sensitivity
Specificity
Metric
Dataset
PERCENTAGE
E
ROC Curves for Test Datasets
GSE106817 AUC = 0.9471
GSE112264 AUC = 0.5728
GSE113486 AUC = 1.0000
GSE113740 AUC = 0.8383
GSE122497 AUC = 0.9680
GSE137140 AUC = 0.9546
GSE139031 AUC = 0.7204
GSE164174 AUC = 0.8341
GSE73002 AUC = 0.9605
True Positive Rate
False Positive Rate

Fig. 5

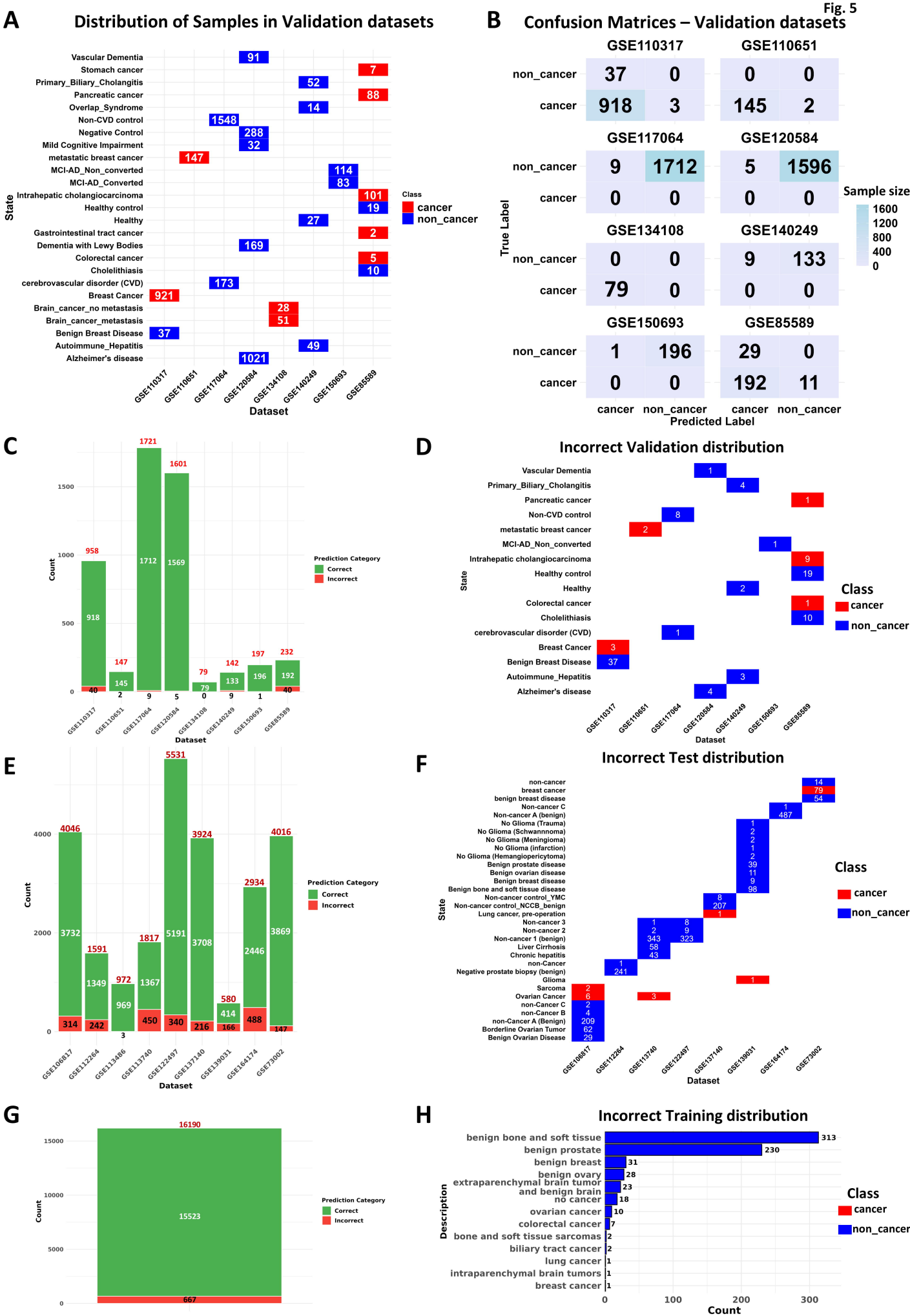
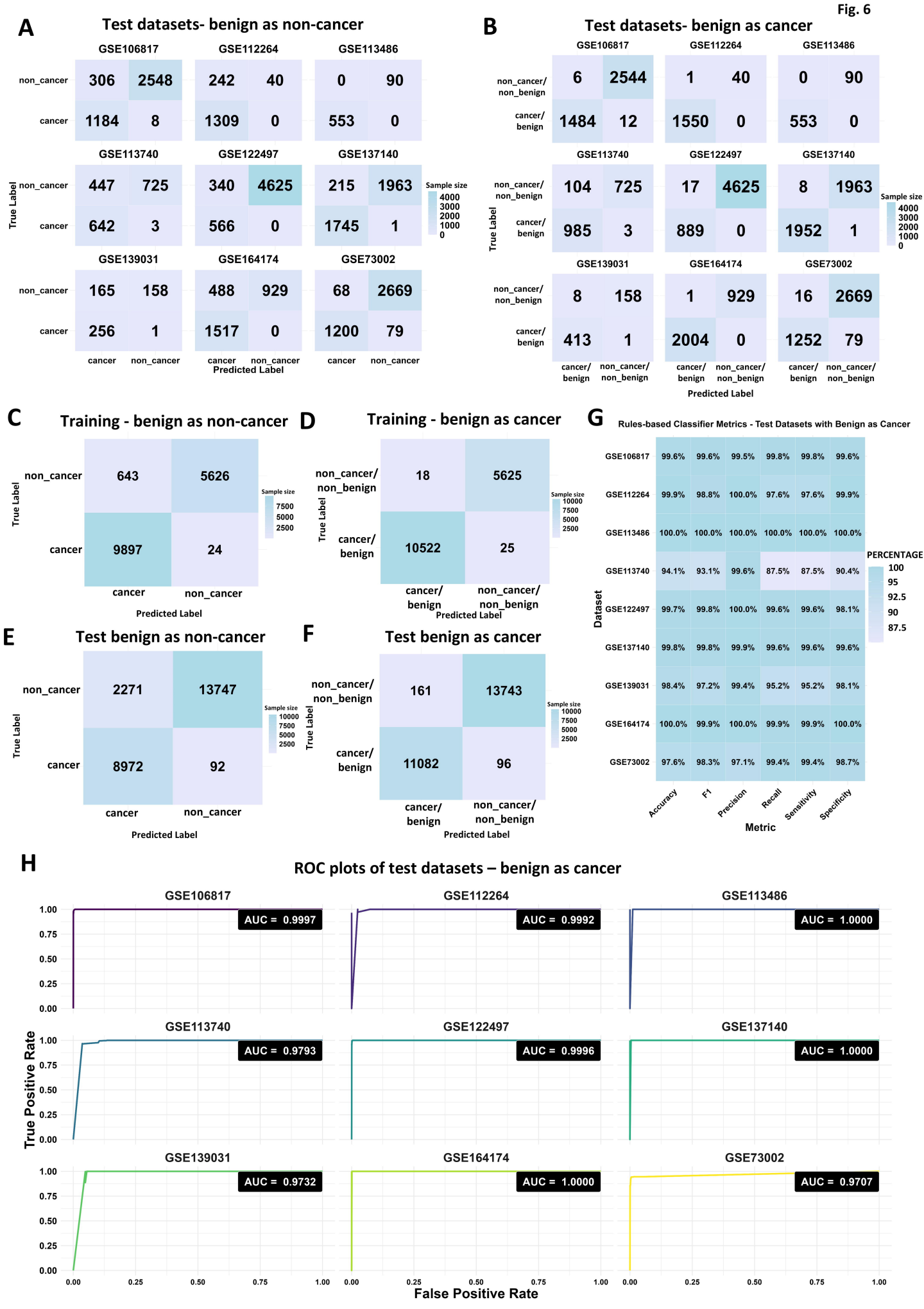

A
Distribution of Samples in Validation datasets
State
Vascular Dementia
Stomach cancer
Primary_Biliary_Cholangitis
Pancreatic cancer
Overlap_Syndrome
Non-CVD control
Negative Control
Mild Cognitive Impairment
metastatic breast cancer
MCI-AD_Non_converted
MCI-AD_Converted
Intrahepatic cholangiocarcinoma
Healthy control
Healthy
Gastrointestinal tract cancer
Dementia with Lewy Bodies
Colorectal cancer
Cholelithiasis
cerebrovascular disorder (CVD)
Breast Cancer
Brain_cancer_no metastasis
Brain_cancer_metastasis
Benign Breast Disease
Autoimmune_Hepatitis
Alzheimer's disease
GSE110317
GSE110651
GSE117064
GSE120584
GSE134108
GSE140249
GSE150693
GSE85589
Dataset
Class
cancer
non_cancer
B
Confusion Matrices – Validation datasets
GSE110317
non_cancer 37 0
cancer 918 3
GSE110651
non_cancer 0 0
cancer 145 2
GSE117064
non_cancer 9 1712
cancer 0 0
GSE120584
non_cancer 5 1596
cancer 0 0
GSE134108
non_cancer 0 0
cancer 79 0
GSE140249
non_cancer 9 133
cancer 0 0
GSE150693
non_cancer 1 196
cancer 0 0
GSE85589
non_cancer 29 0
cancer 192 11
True Label
cancer non_cancer cancer non_cancer
Predicted Label
Sample size
1600
1200
800
400
0
C
Count
Prediction Category
Correct
Incorrect
Dataset
D
Incorrect Validation distribution
State
Vascular Dementia
Primary_Biliary_Cholangitis
Pancreatic cancer
Non-CVD control
metastatic breast cancer
MCI-AD_Non_converted
Intrahepatic cholangiocarcinoma
Healthy control
Healthy
Colorectal cancer
Cholelithiasis
cerebrovascular disorder (CVD)
Breast Cancer
Benign Breast Disease
Autoimmune_Hepatitis
Alzheimer's disease
Dataset
Class
cancer
non_cancer
E
Count
Prediction Category
Correct
Incorrect
Dataset
F
Incorrect Test distribution
State
non-cancer
breast cancer
benign breast disease
Non-cancer C
Non-cancer A (benign)
No Glioma (Trauma)
No Glioma (Schwannnoma)
No Glioma (Meningioma)
No Glioma (infarction)
No Glioma (Hemangiopericytoma)
Benign prostate disease
Benign ovarian disease
Benign breast disease
Benign bone and soft tissue disease
Non-cancer control_YMC
Non-cancer control_NCCB_benign
Lung cancer, pre-operation
Non-cancer 3
Non-cancer 2
Non-cancer 1 (benign)
Liver Cirrhosis
Chronic hepatitis
non-Cancer
Negative prostate biopsy (benign)
Glioma
Sarcoma
Ovarian Cancer
non-Cancer C
non-Cancer B
non-Cancer A (Benign)
Borderline Ovarian Tumor
Benign Ovarian Disease
Dataset
Class
cancer
non_cancer
G
Count
Prediction Category
Correct
Incorrect
H
Incorrect Training distribution
Description
benign bone and soft tissue
benign prostate
benign breast
benign ovary
extraparenchymal brain tumor and benign brain
no cancer
ovarian cancer
colorectal cancer
bone and soft tissue sarcomas
biliary tract cancer
lung cancer
intraparenchymal brain tumors
breast cancer
Count
Class
cancer
non_cancer

Fig. 6


A
Test datasets- benign as non-cancer
GSE106817
non_cancer 306 2548
cancer 1184 8
GSE112264
242 40
1309 0
GSE113486
0 90
553 0
GSE113740
non_cancer 447 725
cancer 642 3
GSE122497
340 4625
566 0
GSE137140
215 1963
1745 1
GSE139031
non_cancer 165 158
cancer 256 1
GSE164174
488 929
1517 0
GSE73002
68 2669
1200 79
cancer non_cancer
Predicted Label
True Label
Sample size
4000
3000
2000
1000
0
B
Test datasets- benign as cancer
GSE106817
non_cancer/ non_benign 6 2544
cancer/ benign 1484 12
GSE112264
1 40
1550 0
GSE113486
0 90
553 0
GSE113740
non_cancer/ non_benign 104 725
cancer/ benign 985 3
GSE122497
17 4625
889 0
GSE137140
8 1963
1952 1
GSE139031
non_cancer/ non_benign 8 158
cancer/ benign 413 1
GSE164174
1 929
2004 0
GSE73002
16 2669
1252 79
cancer/ benign non_cancer/ non_benign
Predicted Label
True Label
Sample size
4000
3000
2000
1000
0
C
Training - benign as non-cancer
non_cancer 643 5626
cancer 9897 24
cancer non_cancer
Predicted Label
True Label
Sample size
7500
5000
2500
D
Training - benign as cancer
non_cancer/ non_benign 18 5625
cancer/ benign 10522 25
cancer/ benign non_cancer/ non_benign
Predicted Label
True Label
Sample size
10000
7500
5000
2500
E
Test benign as non-cancer
non_cancer 2271 13747
cancer 8972 92
cancer non_cancer
Predicted Label
True Label
Sample size
10000
7500
5000
2500
F
Test benign as cancer
non_cancer/ non_benign 161 13743
cancer/ benign 11082 96
cancer/ benign non_cancer/ non_benign
Predicted Label
True Label
Sample size
10000
7500
5000
2500
G
Rules-based Classifier Metrics - Test Datasets with Benign as Cancer
GSE106817 99.6% 99.6% 99.5% 99.8% 99.8% 99.6%
GSE112264 99.9% 98.8% 100.0% 97.6% 97.6% 99.9%
GSE113486 100.0% 100.0% 100.0% 100.0% 100.0% 100.0%
GSE113740 94.1% 93.1% 99.6% 87.5% 87.5% 90.4%
GSE122497 99.7% 99.8% 100.0% 99.6% 99.6% 98.1%
GSE137140 99.8% 99.8% 99.9% 99.6% 99.6% 99.6%
GSE139031 98.4% 97.2% 99.4% 95.2% 95.2% 98.1%
GSE164174 100.0% 99.9% 100.0% 99.9% 99.9% 100.0%
GSE73002 97.6% 98.3% 97.1% 99.4% 99.4% 98.7%
Accuracy F1 Precision Recall Sensitivity Specificity
Metric
Dataset
PERCENTAGE
100
95
92.5
90
87.5
H
ROC plots of test datasets – benign as cancer
GSE106817
AUC = 0.9997
GSE112264
AUC = 0.9992
GSE113486
AUC = 1.0000
GSE113740
AUC = 0.9793
GSE122497
AUC = 0.9996
GSE137140
AUC = 1.0000
GSE139031
AUC = 0.9732
GSE164174
AUC = 1.0000
GSE73002
AUC = 0.9707
True Positive Rate
False Positive Rate
0.00
0.25
0.50
0.75
1.00