Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

PanClassif: Improving pan cancer classification of single cell RNA-seq using machine learning

Kazi Ferdous Mahin*, Md. Robiuddin*, Mujahidul Islam, Shayed Ashraf, Farjana Yeasmin and Swakkhar Shatabda*

Department of Computer Science and Engineering, United International University, Madani Avenue, Satarkul, Badda, Dhaka 1212, Bangladesh

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Cancer is one of the major causes of human death per year. In recent years, cancer identification and classification using machine learning have gained momentum due to the availability of high throughput sequencing data. Using RNA-seq, cancer research is blooming day by day and new insights of cancer and related treatments are coming into light.

Results: In this paper, we propose PanClassif, a method that requires a very few and effective genes to detect cancer from RNA-seq data and is able to provide performance gain in several wide range machine learning classifiers. We have taken 22 types of cancer samples from The Cancer Genome Atlas (TCGA) having 8287 cancer samples and 680 normal samples. Firstly, PanClassif uses k-Nearest Neighbor (k-NN) smoothing to smooth the samples to handle noise in the data. Then effective genes are selected by Anova based test. For balancing the train data, PanClassif applies an oversampling method, SMOTE. We have performed comprehensive experiments on the datasets using several classification algorithms. Experimental results shows that PanClassif outperform existing state-of-the-art methods available and shows consistent performance for two single cell RNA-seq datasets taken from Gene Expression Omnibus (GEO). PanClassif improves performances of a wide variety of classifiers for both binary cancer prediction and multi-class cancer classification.

Availability: PanClassif is available as a python package (https://pypi.org/project/panclassif/)

Contact: mrrobi040@gmail.com, mahin.kkgb.km@gmail.com

Supplementary information: Supplementary data are available at online.

1 Introduction

Cancer remains one of the most spiteful diseases that tribulates human life. It comprises a convoluted biological system that requires meticulous and all-inclusive analysis. Over the last couple of decades, RNA sequencing has become a remarkable way for transcriptome profiling. The revolution sparkled when high throughput sequencing evolved from bulk RNA sequencing to single molecular, single cell (Margulies *et al.*, 2005). Spatial transcriptome analysis methods have enabled increasingly meticulous cell resolution incorporated with spatial information (Hong *et al.*, 2020).

Heterogeneity of cells which comes from different kind of gene regulation, gene mutations and stochastic dissimilarity are reflected at the transcriptomic, proteomic and genomic levels. This kind of heterogeneity is a significant factor in cancer treatment failure. Cancer is itself a complex disease but during formation of malignant cells, the lineages divide and form intra-tumour heterogeneity (Saadatpour *et al.*, 2015). It is clear that conventional bulk detection method which measures the average profile of the tumour population may have limitations to characterize intricate diseases like cancer. In the case of single cell RNA-seq experiments, challenges in measuring technologies like amplification bias, library size differences, cell cycle and low capture rate of RNA lead to substantial noise in which is known as dropouts (Eraslan *et al.*, 2019). It necessitates different smoothing techniques to be applied to the data.

In cancer classification, it is crucial to identify features explicitly because of the high dimensionality of the data. In recent years, machine learning based methods have gained popularity in cancer cell classification due to the effectiveness in identifying crucial features and availability of data from high throughput machines (Kourou *et al.*, 2015).

Li *et al.* (2017) have used RNA-seq expression of 9096 tumor samples which represent 31 types of tumor data collected from TCGA. They have used genetic algorithms for gene selection, log_2-transform for normalizing data and k-NN for classification. After selecting 20 genes, they have performed a classification experiment with those data sets. They

© The Author 2021. All rights reserved.

have also found several genes that are responsible for showing sexual-dimorphism in liver cancer. Hsu and Si (2018) takes 33 types of cancer from TCGA and does a comparison of different machine learning models based on prediction accuracy, recall, precision, F-score, and training time. The models they have chosen are Decision Tree, k-NN, $linear$ and $poly$ kernel Support Vector Machine (SVM), and Neural Networks (NN).

Kim *et al.* (2019) have classified 21 types of cancer using single-cell RNA-Seq gene count data which is collected from TCGA and then filter out genes from each cancer based on the highest F-score. They have chosen n-numbers of genes for training features and figures out the best n genes which gives the best performance. After that they have used z-score for normalizing the data and perform binary and pan-cancer classification using machine learning algorithms. They have used breast cancer and skin melanoma cancer dataset from Gene Expression Omnibus (GEO) repository to validate the overall performance of their model.

Wang *et al.* (2019) have used RNA-seq gene count data and filter out genes that have zero counts in more than 80% of samples. Then they have augmented the data to eradicate the effect of noise in training. Their augmented approach that has made data scale-invariant and finally classify data with the neural network. Guo *et al.* (2020) have discussed a bout molecular heterogeneities and complex micro-environments which bring great challenges for diagnosis of cancer as well as treatment procedure with the help of RNA Seq. Focusing on the process and analysis of single cell RNA-seq data for cancer research, they have developed an R package called *scCancer* which has integrated the basic steps of processing single cell RNA-seq data. It has used common batch-effect correction methods for performing multi-sample integration analysis. They propose a new algorithm named 'NormalMNN'. It can automatically generate user-friendly graphic reports to overview all the results. The experimental dataset that they used for testing has 433405 cells from 56 samples.

Single-cell RNA Seq is the next-generation technology that is able to detect cancer more accurately. However, the identification of different cancer cells from a multi-cancer data sample is still challenging. Single cell RNA-Seq datasets comes with a huge number of features or gene expressions and as well as with noise. As these datasets have high dimensionality, machine learning models often take too much time and resource to train and perform prediction. In addition, there is an imbalance in the datasets. Often the ratio of normal and cancer cell is not balanced as shown in Table 1 and Table 2. Training with imbalanced data leads to biased models with lower sensitivity.

In this paper, we propose PanClassif to address all the above mentioned issues. PanClassif requires a very small amount of genes to correctly classify different types of cancer samples with a very few miss-classifications. We have taken 22 types of cancer data from TCGA. After that we have used smoothing on the datasets using k-NN smoothing and then performed ANOVA based gene selection. After picking the best genes we filter out rest of the genes from the dataset. We then perform oversampling to handle the data imbalance problem on the training data. Then we evaluate the performance of the method using several machine learning models. Most of the models perform very well. We have chosen 6 classifiers: SVM (*linear* kernel), SVM (*rbf* kernel), Random Forest, Neural Network, k-Nearest Neighbour and Adaboost algorithm. Most of the classification algorithms have shown increase in performance following PanClassif methodology. We further performed gene set enrichment analysis over the best selected genes from TCGA cancer datasets and most of the selected genes have shown correlation with cancer/tumor type diseases. Experiments show the effectiveness of the method on Gene Expression Omnibus (GEO) single-cell RNA-Seq datasets. The results obtained by PanClassif are significantly improved over existing state-of-the-art methods.

2 Materials and Methods

In Fig. 1, we have shown the complete workflow of PanClassif. At first, we have downloaded our desired data from TCGA and GEO and then PanClassif applies k-NN smoothing on our downloaded data. Then feature selection is performed to select an effective set of features followed by the standardization of the data. After standardization, PanClassif splits the data in train and test. Oversampling is performed on the train data to address the imbalance issue. Finally, PanClassif trains the data for binary and multiclass classification models using training data and validate the model's performance using test data. In this paper, two classification problems are addressed: binary and multi-class. In binary classification, we checked if a sample has cancer or not and in multi-class classification, we checked which type of cancer a sample has. The detailed discussion about each component of PanClassif are described in the rest of the section.

2.1 Data Collection

In this paper, we have used RNA-seq expression count data of cancerous and their respective normal for this research. We have collected data from the repositories of The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO). From TCGA, we have collected samples of 22 types of cancer and from GEO, we have collected breast cancer (GSE75688) and skin melanoma cancer (GSE72056) samples. The expression counts were calculated using RSEM (Li and Dewey, 2011) on the data downloaded using TCGA2STAT (Wan *et al.*, 2016). TCGA2STAT is also used to download TCGA data. Total 8287 cancer samples and 680 normal samples are collected from TCGA having 20,501 genes. Breast cancer (BRCA) and skin melanoma cancer (SKCM) data that are collected from GEO have 317 cancer samples, 198 normal samples and 1257 cancer samples, 3256 normal samples respectively. The summary of the datasets collected are given in Table 1 and Table 2. Both of the datasets were then prepared for binary cancer prediction and multiclass cancer type identification problems. Single cell RNA-seq data from GEO datasets were used to validate the effectiveness of the methodology.

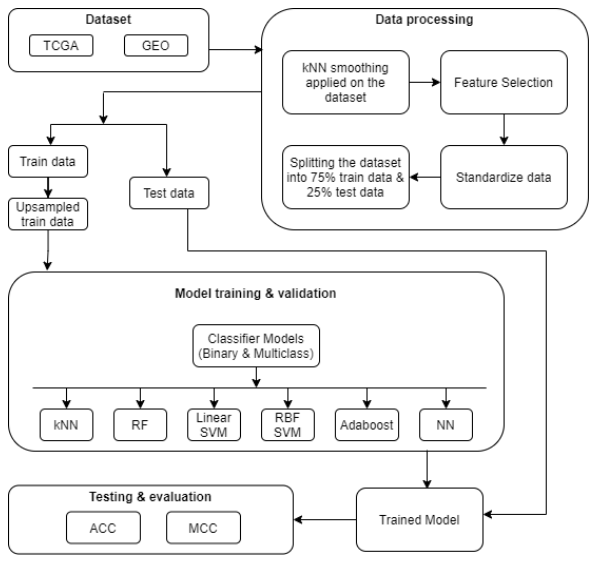


Fig. 1. Complete workflow of PanClassif.

Table 1. Breast cancer and skin melanoma cancer datasets from GEO.

Cancer Type	Accession code	Cancer	Normal
Breast Cancer	GSE75688	317	198
Skin Melanoma	GSE72056	1257	3256
Total		1574	3454

Table 2. TCGA cancer codes and their respective cancer and normal samples used in our work.

Code	Cancer	Normal	Code	Cancer	Normal
BLCA	427	19	LUAD	576	58
BRCA	1212	112	LUSC	552	51
CESC	309	3	PAAD	183	4
CHOL	45	9	PCPG	187	3
COAD	328	26	PRAD	550	52
ESCA	198	11	READ	105	6
HNSC	566	43	SARC	265	3
KICH	91	25	STAD	450	32
KIRC	606	72	THCA	568	59
KIRP	323	32	THYM	122	3
LIHC	423	50	UCEC	201	7
Total				8287	680

2.2 Data Smoothing

We have already mentioned that single-cell RNA-seq data tends to be extremely noisy (Wagner *et al.*, 2018) . In order to remove the effect of noise and variability of the data, we have used k-nearest neighbour smoothing. The k-NN smoothing uses step-wise approach to identify the nearest neighbours and Freeman-Tukey transformation to ensure that in the distance calculation, all genes expression levels contribution are approximately equal. In addition to this, Principal Component Analysis (PCA) is also used to remove noise and give attention to the most important expression differences among cells. We have performed this smoothing on our data samples using $k = 10$. Dataset having less than 10 samples is skipped in this process.

2.3 Feature Selection

PanClassif uses $F1$-Anova test to select effective set of genes from the huge number of genes for which the expression data is available. The feature selection algorithm in depicted in Algorithm 1. For each cancer type, first the dataset for the particular cancer is extracted from database as gene expression count. Then Anova test is performed to get the scores. Top n genes are selected and only the unique genes are added to the selected gene list. We have performed detailed experiments to show the effectiveness of this feature selection algorithm and the effect of the parameter n. The algorithm is initially experimented on TCGA dataset and then validated using the same parameters in the GEO dataset.

2.4 Standardization and Sampling

Feature scaling or standardization is a very important step in data pre-processing. It helps to reduce effect of any gene that shows high expression count and can effect the classification model performances. PanClassif uses z-score standardization on the datasets. Standard score of a sample x_{ij} drawn from a population j with mean, μ can be calculated using:

$$z_{ij} = \frac{(x_{ij} - \mu_j)}{\sigma_j}, \forall_i \forall j \tag{1}$$

Here, σ_j is the standard deviation of the population . Here, μ_j, σ_j are estimated using the sample mean and standard deviation.

Algorithm 1: Feature Selection

Input: The cancer type set $C = \{c_1, c_2, \ldots, c_l\}$, the cancer gene expression database $CancerDB$ and a positive number n which is a parameter to select genes

Output: A list of selected genes, $G = \{g_1, g_2, \ldots, g_k\}$

```
G ← ∅
foreach c_i ∈ C do
    geneExp ← GETCANCERDATA(CancerDB, c_i)
    geneScores ← PERFORMANOVA(geneExp)
    geneList ← SELECTTOPGENES(geneScores, n)
    foreach g ∈ geneList do
        if g ∉ G then
            G = G ∪ {g}
return G
```

After standardization, data was randomly split into train and test datasets with a 3:1 ratio. We have used this ratio as done by Kim *et al.* (2019). In both of the datasets, we could notice that there is an imbalance in the number of positive and negative classification. particularly for the TCGA datasets, the number of normal or non-cancer data is very low. In case of binary classification the ratio is 12:1. The GEO dataset is also imbalanced in the other way around. Such imbalance creates a bias in the learned model. To address this problem, PanClassif uses Synthetic Minority Oversampling Technique (SMOTE) (Chawla *et al.*, 2002) to oversample the train dataset by generating synthetic data in the minority class.

2.5 Machine Learning models for classification

The selection of proper classification model is a very crucial step in machine learning. However, to show the effectiveness of PanClassif, we have selected a number of existing classifiers to apply them for binary and multi-class classification tasks. They are k-Nearest Neighbour (k-NN), Support Vector Machines with linear kernel (L-SVM), Support Vector Machines with radial basis function kernel (RBF-SVM) (Cortes and Vapnik, 1995), Random Forest (RF) (Tin Kam Ho, 1998), Adaboost (Freund and Schapire, 1997) and Artificial Neural Networks (NN) (Hopfield, 1982). In k-NN, we have used $k = \{1, 2, 3, 5, 10, 13, 17\}$. For RF, the *maxDepth* parameter of decision trees were kept 100. For both of L-SVM and RBF-SVM, C was kept 1 and γ was set at the inverse value of multiplication of no. of the features in the dataset and their variance. Decision tree is used as base classifier for Adaboost. For neural network, we have used 5 hidden layers apart from the input layer and output layer. The number of neurons in the hidden layers were 512, 256, 128, 64, and 32 respectively. Same hyper-parameters are used in both binary and multi-class classification tasks except in the neural network. In neural network, sigmoid is used as last layer activation for binary and softmax is used as last layer activation in multi-class classification.

2.6 Performance Evaluation

For evaluating the performance of the classification models and overall PanClassif methodology we have used two metrics: Accuracy (ACC) and Mathews Correlation Coefficient (MCC). In the case of binary classification task, true positives (TP) is the number of positive class instances classified correctly and true negatives (TN) are the number of negative class instances classified correctly by the model. On the other had false positives (FP) and false negatives (FN) are number of negative and positive class instances respectively that are wrongly classified in the

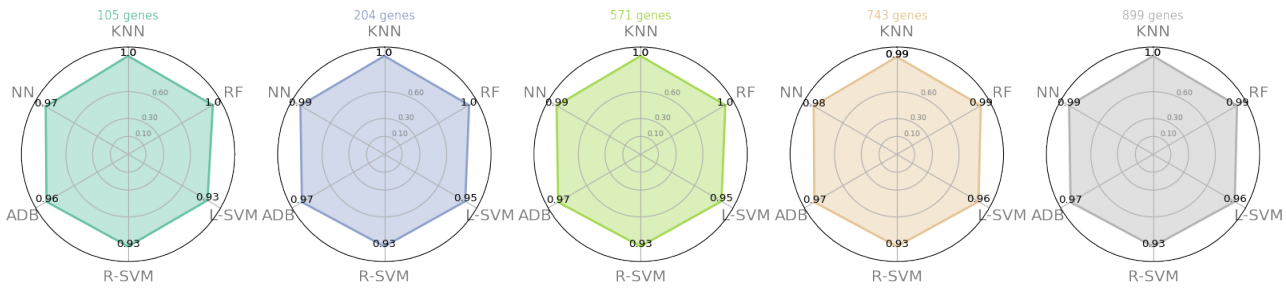


Fig. 2. Spider plot for ACC score for the binary classification task in TCGA data using 5 different datasets created selecting different number of genes.

opposite group. Based on these, accuracy and MCC are defined as in the following.

Accuracy: Accuracy determines how accurately the model classifies the instances. It has a range of 0 to 1 where a value 1 indicates the best possible classifier. However, please note that for imbalanced data, accuracy might not always reflect the proper model performances. Accuracy is define in the following equation:

$$ACC = \frac{TN + TP}{TP + TN + FP + FN} \tag{2}$$

Mathews Correlation Coefficient: Mathews Correlation Coefficient or MCC provides a good performance measure even when the data is imbalanced. It has a range of -1 to +1. A perfect model will show a value close to +1 and a less accurate model will achieve a value less than that and even negative.

$$MCC = \frac{(TN * TP) - (FN * FP)}{\sqrt{(TN + FP)(TP + FP)(TN + FN)(TP + FN)}} \tag{3}$$

Calculating accuracy and MCC score for multi-class is different than binary. In the case of multi-class task, we have used a *one vs all* for each class and then calculated accuracy and MCC score for each class. After that, the weighted average value of each metric is reported.

3 Results and Discussion

All the experiments done in this paper and the reported results are prepared using a machine having 32GB RAM, Xenon processor, and Linux operating system. All the programs are written in Python 3.8 and Scikit-learn library. Each experiments were run 5 times and only the average results are reported. Please note that Supplementary data contains more detailed results on the datasets and methods used in this paper.

3.1 Performance on Binary Classification Task

We have used 75% of the data for training and the rest of the data for test purpose. In case of binary classification task with TCGA data, the training data has a very high imbalance ratio of 12:1 with 6213 cancer samples and 510 normal samples. The oversampling techniques reduced this ratio to 3:1 resulting into 2057 normal samples. We have annotated two classes, 'cancer' samples as class 1 and 'normal' samples as class 0. Then we have prepared 5 different datasets with top 105, 204, 571, 743, 899 genes (keeping the parameter in feature selection $n = 5, 10, 30, 40, 50$ respectively) and train models with those 5 datasets and test with the remaining 25% test data. A spider plot of accuracy score achieved by different classifiers on different number of selected genes are shown in Fig. 2. From the results shown in Fig. 2, we observe that k-NN and Random Forest (RF) are the two best performing classifiers. Adaboost (ADB) and the variants of SVM did not perform well in any of the datasets compared to them. However, performance of Neural Network was very close to the best results where number of selected genes were 204, 571 and 899. We have also validated these results on the single cell RNA-seq GEO datasets. The settings and parameters of PanClassif pipeline were kept same and similar pattern is observed in the results (see Supplementary data for details tables and graphs). Please note that, results in terms of MCC also shows superiority of Random Forest and k-NN over other classifiers.

3.2 Performance on Pan Cancer Classification

For pan cancer classification, we create 5 datasets with top 105, 204, 571, 743, 899 genes in the similar way and annotate them with 23 classes where 22 types were of cancer class and 1 normal class. After that, we have trained all 6 Machine Learning models with the same configurations described in Section 2.5. The train and test set ratio was kept similar as in the binary prediction task (3:1). A spider plot of accuracy scores achieved by the different classifiers using different number of genes is shown in Fig. 3. The plot shows a very poor performance by the AdaBoost algorithm. The SVM classifier with radial basis function as kernel showed some satisfactory results. However, the rest of the classifiers performed very well in terms of accuracy (results in terms of MCC, please see Supplementary data). Among all the classifiers, Random forest performed the best and the confusion matrix for Random forest in the multi-class classification is shown in Fig. 4. From this matrix, it is observable that there were no inter-class mistakes done by the classifier. However, a few false positives and false negatives were found for cancer types BRCA, BLCA, KICH, LIHC, KIRP, THCA.

To validate the results, a similar set of experiments were performed on single cell RNA-seq GEO datasets to. Those results also show similar patterns (please check supplementary results).

3.3 Comparison with other methods

We have also performed a comparative analysis for the performance of PanClassif with that of the work in Kim *et al.* (2019), which is a state-of-the-art method for cancer classification and recognition task from single cell RAN-Seq data. Table 3 shows the MCC and accuracy achieved by these methods on TCGA data for both of the tasks using different classifiers. The results reported in the table were taken from the reported results by Kim *et al.* (2019). For binary classification task, Kim *et al.* (2019) achieved highest accuracy of 0.99 with 300 genes using neural network. However, using PanClassif, accuracy of neural network is 0.99 using 204 genes. Our best classifier is random forest which gives 1.00 accuracy with 105 genes. In pan-cancer classification the performance of PanClassif remain same. However, for pan cancer classification, Kim *et al.* (2019) achieves a best accuracy of 0.90 by neural network using 300

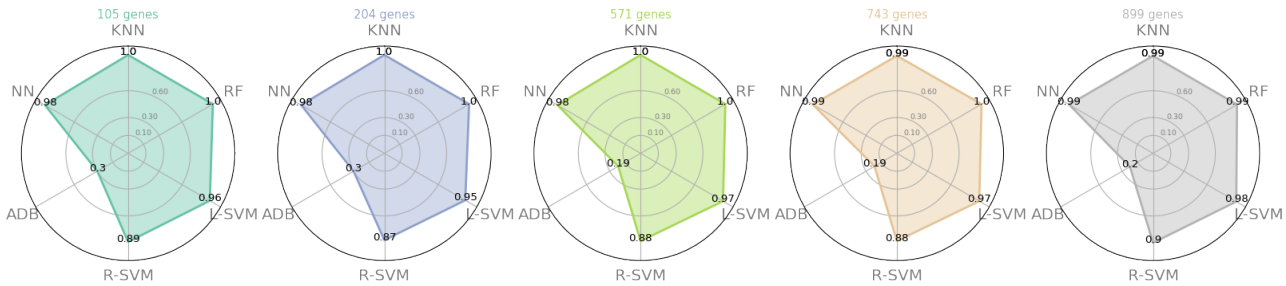


Fig. 3. Spider plot for ACC score for the multi-class classification task in TCGA data using 5 different datasets created selecting different number of genes.

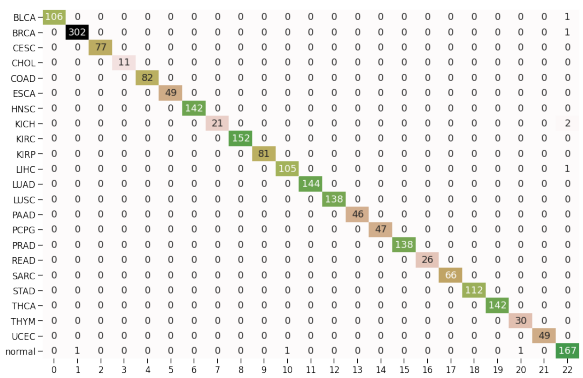


Fig. 4. Confusion matrix of multi-class classification by Random Forest Classifier.

genes. PanClassif achieves 0.98 accuracy using neural network with 105 genes. Kim *et al.* (2019) achieved best performances using in between 200 to 300 genes for binary and pan-cancer classification. In our case, the range of genes is 105-204 in most cases except SVM with *rbf* kernel shows top-notch performance. In case of MCC score, PanClassif also significantly outperform state-of-the-art method.

In case of the GEO datasets, Kim *et al.* (2019) achieves best MCC of 0.85 for binary classification and 0.86 in multiclass classification as reported in their paper. Their best model is ANN and 0.85 and 0.86 are the MCC score of ANN classifier. For PanClassif, ANN got 0.97 MCC score in binary and 0.93 in multiclass classification. Except for SVM with *linear* kernel, all of our classifiers using PanClassif get significantly higher MCC score than their best reported model.

	binary				**multi-class**			
	Panclassif		Kim *et al.* (2019)		Panclassif		Kim *et al.* (2019)	
Models	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC
kNN	0.97	1	0.8	0.92	0.995	1	0.71	0.72
RF	0.97	1	0.83	0.95	0.996	1	0.86	0.86
L-SVM	0.72	0.95	0.69	0.90	0.96	0.96	0.87	0.88
Rbf-SVM	0.67	0.93	0.83	0.95	0.88	0.89	0.87	0.86
ANN	0.94	0.99	0.92	0.99	0.98	0.98	0.89	0.90

Table 3. Performance comparison of PanClassif with Kim *et al.* (2019) on TCGA dataset.

3.4 Comparative Analysis

In this section, we provide experimental analysis on the strengths of PanClassif. Firstly, to show the effectiveness of the Anova based feature selection, we have compared it with χ^2 based feature selection and compared the results with Anova based method. Average MCC and accuracy for each classifier were taken using the best 105, 204, 571, 743, 899 genes selected by both methods. A box plot of the MCC for these two methods for binary and multi-class classification task is shown in Fig. 5(a). As shown in the figure, Anova based feature selection performs significantly better MCC scores for both tasks. ANOVA achieves average MCC of 0.84 and 0.85 compared to those of 0.80 and 0.83 achieved by χ^2 method.

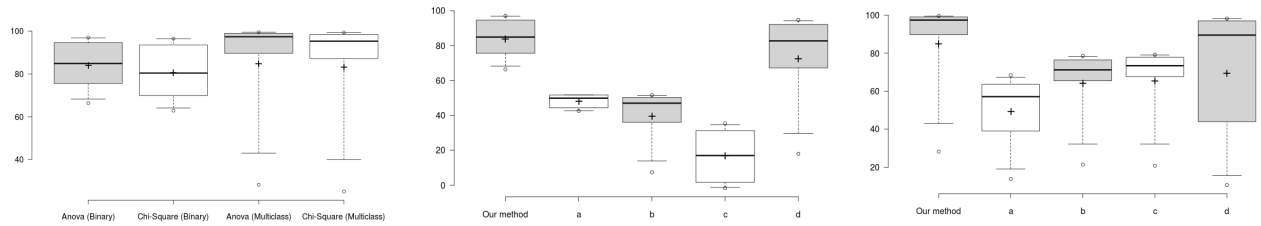
To show the effectiveness of smoothing and oversampling steps of PanClassif, we have created 4 variants: (a) under-sampled data with smoothing, (b) over-sampled data without smoothing and, (c) imbalanced data without smoothing (d) imbalanced data with smoothing. Fig. 5 (b) and (c) shown box plot of MCC scores for all these variants along with PanClassif run using different models on binary and multi-class classification tasks on TCGA data (please see supplementary data for deails).

In the case of undersampling, our train data with 6213 cancer samples were down-sampled by picking 8.20% samples randomly from each cancer types resulting into 509 cancer samples. This is variant 'a' in Fig. 5 (b-c). This variant achieves average MCC score of 0.48 and 0.49 in binary and multi-class tasks which is very poor compared to those of 0.84 and 0.85 respectively. The other variant 'b' in Fig. 5 (b-c) was created by oversampling and removing smoothing step. This variant shows the superiority of oversampling over undersampling in multi-class classification by giving MCC values of 0.39 and 0.64 in binary and multi-class respectively.

The variant 'c' in Fig. 5 (b-c) was generated without over-sampling and without smoothing which shows an average MCC of 0.16 in binary and 0.65 in multi-class classification. However, adding smoothing in imbalance data, dictates its excellence by an average MCC of 0.77 and 0.69 in binary and multi-class classification respectively. This case is represented by the variant 'd' in Fig. 5 (b-c).

3.5 Gene Enrichment Analysis

We have further performed gene enrichment analysis with our top 105 genes using the 'Enrichr web tool' (Kuleshov *et al.*, 2016; Chen *et al.*, 2013) to validate if the top genes found by PanClassif are associated with cancer cell line. A horizontal bar plot of our top genes against the different cell lines are shown in Fig. 6. The figure shows 6 cancer enriched genes among top 10. They are Pulmonary Sclerosing Hemangioma, Clear-cell metastatic renal cell carcinoma, Conventional (Clear Cell) Renal Cell carcinoma, Invasive Ductal breast Carcinoma, Pancreatic Neoplasm and Kidney Neoplasm'. With this finding, we can conclude that PanClassif is able to identify cancer related best significant genes and that is why all classifiers are performing well.



(a) (b) (c)

Fig. 5. Box plot of MCC scores for (a) different feature selection techniques in binary and multi-class classification tasks, different components of PanClassif on (b) binary classification task and (c) multi-class classification task.

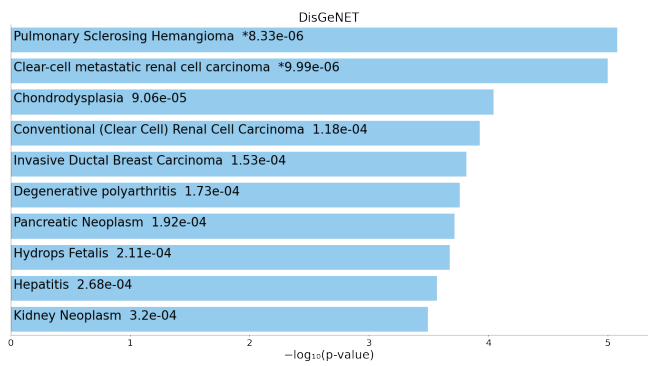


Fig. 6. Bar Chart of gene enrichment analysis.

3.6 Gene Significance Test

To further test the effectiveness of PanClassif, we have performed gene significance test in two separate experiments. In the first experiment, we have created two datasets: raw data from the TCGA repository and the other one after applying smooting and over-sampling as defined in PanClassif. This time we have selected only 22 genes by selecting only the top gene from each cancer type. After that we have used a neural network classifier to perform multi-class classification on these two datasets. In Table 4, we have shown best genes selected for each cancer and respective sensitivity or recall for each of the dataset. As we can see in Table 4, except CHOL and READ cancer type, PanClassif improves the sensitivity of classifying a cancer type using the top 22 genes. Also note that two cancer types, BRCA and PCPG have already attained a 100% sensitivity using only 22 genes.

cancer type	top gene	raw data	Pan-Classif	cancer type	top gene	raw data	Pan-Classif
BLCA	CLEC3A	37.25	89.71	LUAD	STX10	40.97	87.5
BRCA	ADAMTS4	66.78	100	LUSC	GDR115	46.61	81.88
CESC	SEH1L	32.47	92.21	PAAD	CT45A1	50.00	78.26
CHOL	PROX2	16.66	0.00	PCPG	FRMPD4	76.92	100
COAD	SCARA3	47.69	86.25	PRAD	TMIGD2	58.74	87.68
ESCA	RGMA	16.98	18.36	READ	LRRC39	11.11	0.00
HNSC	CA9	48.95	88.02	SARC	PTGR2	65.33	98.48
KICH	DZIP1L	42.11	56.52	STAD	GCNT7	46.55	86.61
KIRC	SFPQ	66.45	84.86	THCA	TLA2G7	73.05	73.94
KIRP	TEKT5	64.36	86.42	THYM	S100B	79.31	56.67
LIHC	ANGPTL5	66.39	83.01	UCEC	TRIM16	39.65	85.71

Table 4. Comparison table of true positive rates between raw data and PanClassif process data using top 22 genes.

In the second experiment, we have used t-SNE plot on the TCGA and GEO datasets. This time we have used 105 top genes which produced satisfactory results in classification. The t-SNE plots for TCGA multi-class classification task is shown in Fig. 7 for raw data and data transformed using PanClassif. The plots on GEO data is reported in the Supplementary file. From the figures, we can clearly see that the samples is clustered according to their classes in the transformed version using PanClassif. On the other hand, in Fig. 7(a) samples of each classes overlaps one another (no transformation). This will lead to errors in classification.

3.7 Availability of PanClassif

PanClassif is available as Python package in Python package index (PyPI). The latest version can be installed using `pip install panclassif`. Documentation of PanClassif is also available at `https://pypi.org/project/panclassif/#description`. PanClassif requires RNA-Seq data as input after smoothing. PanClassif requires a number of folders to be specified as data input and very easy to use.

3.8 Discussion

PanClassif presents a pipeline to treat TCGA pan cancer RNA-seq samples in such a way that different cancer phenotype can be classified and distinguished very efficiently with a less number of features. After applying PanClassif on TCGA datasets using 105 top genes we have found better distinguishable clusters in comparison to untreated data as shown in Fig. 7. PanClassif achieves consistent performance in most of the machine learning models in terms of accuracy and MCC in TCGA and GEO datasets. For Binary Classification of cancer and normal samples data from TCGA, PanClassif is able to classify more than 99% accurately using both k-nearest neighbour and Random Forest with the MCC 0.97 on both models. These results are significantly better than those reported by Kim *et al.* (2019) using 300 genes. However, Adaboost models(decision tree as base classifier) have performed very poorly on TCGA data as the datasets are high dimensional data which might have lead to curse of dimensionality. Due to this reason the calculation complexity of decision tree classifier have increased exponentially with the increasing depth of tree thus ADB perform poorly.

We have further experimented PanClassif with single cell RNA-seq data from Gene Expression Omnibus (GEO) and got the consistent performance. We get MCC 0.98 on k-NN, MCC 0.99 on RF and MCC 0.98 on ANN with the GEO data as well, but L-SVM perform poor with GEO datasets and the reason behind that can be skewed data. L-SVM is not a good choice for skewed datasets.

We have performed gene enrichment analysis with top 105 genes and came to know that 'PTHLH' gene can disable or hamper cell apoptosis which prevent cancer or tumor cells to die. We have found that if we select best one gene from each cancer type of TCGA using PanClassif still

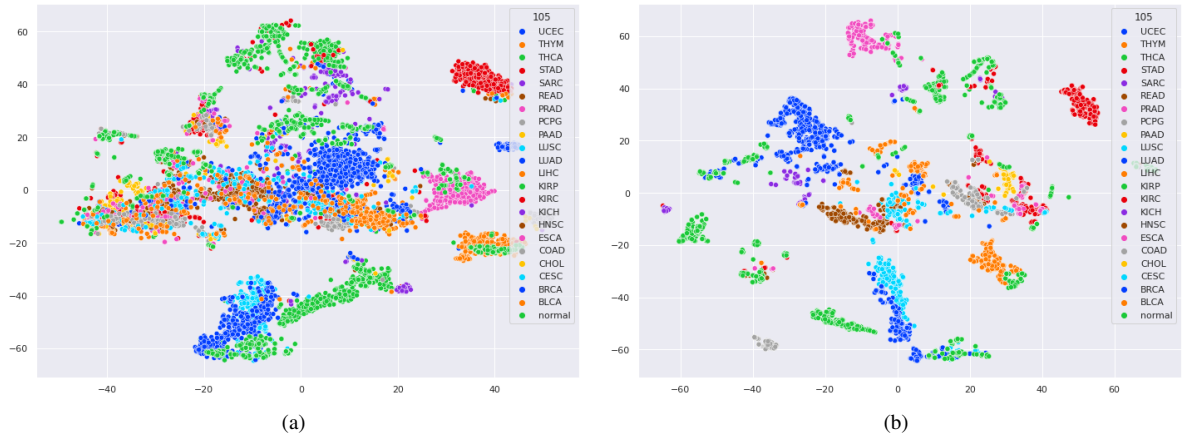


Fig. 7. t-SNE of (a) raw TCGA data with no transformation and (b) TCGA's processed data using PanClassif.

they can correctly classify more than 74% cancer and normal samples. This might be due to the fact that these particular genes are greatly responsible for those cancer phenotype. Further research may reveal there other significance or insight in cancer study. Our method can help cancer study related to pan cancer types.

4 Conclusion

In this paper, we propose PanClassif, an effective single cell RNA-seq data based cancer classification method. Experimental results shows the effectiveness of the proposed method on binary cancer prediction and type identification on TCGA and GEO datasets and thus achieving significantly improved results over state-of-the-art method. The gene enrichment analysis and significance tests shows evidence for the overall effectiveness of PanClassif. Finally, a python package is made available for use. We believe, it is possible to extend the PanClassif methodology to general cell classification tasks using single-cell RNA-Seq data. It is important to note that most of the single cell experiments might not show the real expression counts of the genes and it is possible to further improve the results on a wider range of datasets using sophisticated imputing techniques.

References

Chawla, N. V. *et al.* (2002). Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, **16**(1), 321–357.

Chen, E. Y. *et al.* (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, **14**, 128.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**(3), 273–297.

Eraslan, G. *et al.* (2019). Single-cell rna-seq denoising using a deep count autoencoder. *Nature Communications*, **10**(1), 390.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, **55**(1), 119–139.

Guo, W. *et al.* (2020). scCancer: a package for automated processing of single-cell RNA-seq data in cancer. *Briefings in Bioinformatics*. bbaa127.

Hong, M. *et al.* (2020). Rna sequencing: new technologies and applications in cancer research. *Journal of Hematology & Oncology*, **13**(1), 166.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, **79**(8), 2554–2558.

Hsu, Y. H. and Si, D. (2018). Cancer Type Prediction and Classification Based on RNA-sequencing Data. *Annu Int Conf IEEE Eng Med Biol Soc*, **2018**, 5374–5377.

Kim, B.-H. *et al.* (2019). Cancer classification of single-cell gene expression data by neural network. *Bioinformatics*, **36**(5), 1360–1366.

Kourou, K. *et al.* (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, **13**, 8 – 17.

Kuleshov, M. V. *et al.* (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*, **44**(W1), W90–97.

Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, **12**(1), 1–16.

Li, Y. *et al.* (2017). A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC Genomics*, **18**(1), 508.

Margulies, M. *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057), 376–380.

Saadatpour, A. *et al.* (2015). Single-cell analysis in cancer genomics. *Trends in genetics : TIG*, **31**(10), 576–586. S0168-9525(15)00140-7[PII].

Tin Kam Ho (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(8), 832–844.

Wagner, F. *et al.* (2018). K-nearest neighbor smoothing for high-throughput single-cell rna-seq data. *bioRxiv*.

Wan, Y.-W. *et al.* (2016). Tcga2stat: simple tcga data access for integrated statistical analysis in r. *Bioinformatics*, **32**(6), 952–954.

Wang, T. *et al.* (2019). Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data. *BMC Bioinformatics*, **20**(1), 40.