Neural Computing and Applications (2022) 34:8447–8454
https://doi.org/10.1007/s00521-021-05739-8

S.I. :WORLDCIST'20

(CDRGI)-Cancer detection through relevant genes identification

Feras Al-Obeidat[1] · Álvaro Rocha[2] · Maryam Akram[3] · Saad Razzaq[3] · Fahad Maqbool[3]

Received: 17 September 2020 / Accepted: 16 January 2021 / Published online: 5 February 2021

Abstract
Cancer is a genetic disease that is categorized among the most lethal and belligerent diseases. An early staging of the disease can reduce the high mortality rate associated with cancer. The advancement in high throughput sequencing technology and the implementation of several Machine Learning algorithms have led to significant progress in Oncogenomics over the past few decades. Oncogenomics uses RNA sequencing and gene expression profiling for the identification of cancer-related genes. The high dimensionality of RNA sequencing data makes it a complex and large-scale optimization problem. CDRGI presents a Discrete Filtering technique based on a Binary Artificial Bee Colony coupling Support Vector Machine and a two-stage cascading classifier to identify relevant genes and detect cancer using RNA seq data. The proposed approach has been tested for seven different cancers, including Breast Cancer, Stomach Cancer (STAD), Colon Cancer (COAD), Liver Cancer, Lung Cancer (LUSC), Kidney Cancer (KIRC), and Skin Cancer. The results revealed that the CDRGI performs better for feature reduction while achieving better classification accuracy for STAD, COAD, LUSC and KIRC cancer types.

Keywords Support vector machine · Cascading classifier · Discrete filtering · Artificial bee colony · Gene expression · CatBoost classifier · Convolutional neural network

1 Introduction

Based on the world cancer report published by the World Health Organization (WHO), cancer rates would upsurge by 50% to a total of 15 million new cases in 2020. Cancer is the main cause of death around the globe, every 6th death is because of cancer [1]. In 2018, around 9.6 million peoples died across globe due to different types of cances. This increase is exerting noticeable pressure on communities and health systems at all economic levels [2]. Cancer has proved to be a life-threatening disease. Hence, there is a great demand for novel techniques for preventing, diagnosing, treating, and curing such a lethal disease. Statistics reveal that the second leading cause of death after cardiovascular disease (CVD) is cancer [1]. The most common cancers are stomach (1.03 million), skin (1.04 million), prostate (1.28 million), colorectal (1.80 million), lung, and breast (2.09 million each) cases [1]. Lung, colorectal, stomach, liver, and breast cancers are reported to be more deadly than other cancer types. Cancer is attributed to a related group of diseases. One common characteristic of all cancer types is that cells start dividing uncontrollably and spread into surrounding tissues. In the human genome, various genes are responsible for controlling the growth and division of the cells. It is a genetic disease where changes in these genes cause cancer. The cause of these genetic mutations may be an inheritance from parents or can arise in a person's course of life

✉ Saad Razzaq
saad.razzaq@uos.edu.pk

Feras Al-Obeidat
feras.al-obeidat@zu.ac.ae

Álvaro Rocha
amr@iseg.ulisboa.pt

Maryam Akram
maryam.akram001@gmail.com

Fahad Maqbool
fahad.maqbool@uos.edu.pk

[1] College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates

[2] ISEG Lisbon School of Economics and Management, University of Lisbon, Lisbon, Portugal

[3] Department of Computer Science and IT, University of Sargodha, Sargodha, Pakistan

because of errors in the cell division process. These errors may be caused due to smoking, radiation, diet, or hormonal changes. The Deoxyribo Nucleic Acid (DNA), which is damaged by certain environmental exposures, can also cause cancer. Mutated genes may develop additional mutations in other genes. These mutations collectively may result in cancerous cells [3]. Cancer genomics is a relatively new research area that benefits from recent technological advances for studying the human genome. A scientist can discover genetic differences that cause cancer by comparing sequencing Ribo Nucleic Acid (RNA) and DNA of cancer cells and normal cells. A timely and accurate cancer diagnosis is still the biggest challenge. Therefore auxiliary measures are required besides conventional clinical tests (Magnetic Resonance Imaging, ultrasonography, and Computed Tomography) for accurate and timely diagnosis [4].

The Cancer Genome Atlas (TCGA) [5] generated datasets comprising copy number variation, microRNA expression gene, DNA methylation, protein expression, and somatic mutation. All such advances have led researchers to develop tools and techniques that could help in better diagnosis, treatment, and prevention of cancer. Machine Learning (ML) and Artificial Intelligence (AI) algorithms can be used as diagnostic tools for cancer by analyzing these data sets. These algorithms can detect patterns that are too indistinct for the human eye to recognize and guide oncologists for better and timely diagnosis and treatment. AI is being applied by some researchers for screening tests so that disease can be caught earlier or potential people at cancer risk can be identified.

Information in a gene, the sequence of DNA base pairs, is turned into functional gene products like proteins or RNA using gene expression. In this process, the DNA is first transcribed into RNA, which is then translated into proteins. Gene expression data estimate the amount of activity of genes within a particular tissue, thus enabling us to get information regarding the corresponding cells' complex activities. Gene expression can show abnormalities in cancerous cells.

CDRGI proposes discrete filtering for Feature Selection (FA) followed by Cascading Classifier to classify tumor as a malignant and benign samples using high dimensional RNA-Seq gene expression data. Discrete filtering is perfromed using Binary Artificial Bee Colony and Support Vector Machine (SVM) to select the fittest and relevant subset of features. This selection helps to achieve better classification accuracy. These selected features are passed to a cascading classifier, which comprises two stages. CatBoost classifier is used at the first stage, followed by the second stage containing Random Forest (RF), Multi Layer Perceptron (MLP), and SVM. Majority voting is used to classify the sample at the second stage. These classification models are selected based on their performance evaluation for Cancer detection.

2 Related work

In this section, related studies that applied various ML and Deep Learning techniques to analyze gene expression data are analyzed. Adam Slowik et al. [6] presented a deep learning approach that used convolutional neural network (CNN), Decision Tree (DT), and Binary Particle Swarm Optimization (BPSO) for classification of various cancers. Classification is based on the RNA sequence of gene expression. The proposed method comprises three phases. During the initial phase, pre-processing is performed using BPSO.In this phase, RNA seq data's best features are selected and transformed into a 2D image. The second stage, namely the augmentation phase, increases the data volume 5 times. The last phase employed a deep convolutional neural network that classifies data into five different cancer types. The authors [7] used RNA seq data available at TCGA, and various types of cancers classification. They have applied different DT, K-nearest neighbors (KNN), ANN, and SVM. The results showed that SVM outperformed the other classifiers. Stacked Denoising Autoencoder (STAD) was developed by [8] to reduce relevant features from gene expression profiles. The usefulness of reduced features was determined using supervised classification models. Lastly, STAD connectivity matrices were analyzed to create a highly interactive gene set. The results showed that these genes' set could act as cancer biomarkers that need to be further studied.

Xaio et al. [9] proposed a deep learning method based on auto-encoder (SSAE) for cancer prediction. The presented SSAE-based method was applied on three publicly available RNA seq datasets related to lungs, stomach, and breast cancers. The results showed that the SSAE model attained the finest classification accuracy. Xiao et al. [10] authors presented an ensemble method based on multimodel using deep learning. The informative genes were initially chosen by differential gene expression analysis and passed onto models for classification. The results showed that prediction accuracy increased as compared to the single classifier. Saini et al. [4] presented gene masking for feature reduction to improve classification accuracy. Gene masking was based on a binary encoded genetic algorithm and integrated with a classifier. The technique was applied on publicly available gene expression data sets, thus reducing the number of features while sustaining classification accuracy. Shaheed et al. [11] suggested two-stage gene selection for classification of cancer that used novel hybridization of harmony search and cuckoo search. Initially, maximum relevance and minimum redundancy

are used to create a related subset of gene results that have shown competitive results compared with related algorithms.

Lu et al. [12] implemented the FA method that used adaptive GA and mutual information maximization (MIM) to classify gene expression datasets. Adaptive GA brought modification in conventional Genetic Algorithm (GA) by introducing the concept of crossover and mutation probabilities. The results have shown the effectiveness of the MIM-AGA over other FA algorithms. One of the vital pathological steps in prostate cancer diagnosis and treatment is to estimate tumor location. Location refers to the laterality of the tumor (can be unilateral or bilateral). Thus, authors in [13] presented machine learning techniques capable of detecting potential biomarkers(genes)at three different prostate locations with high accuracy. Shon et al. [14] suggested a cost-sensitive hybrid deep learning approach. This approach extracted genes that are significant for prognosis prediction of kidney cancer. It employed a combination of a deep symmetric autoencoder and neural network. In 5 layered DAE model first two layers are for encoding, the third layer is for gene extraction, and the last two layers perform decoding. In CDRGI our focus is to detect cancer of seven types by using minimum features set. We have used high dimensional RNA-sequential gene expression data. Discrete filtering is performed to select the most relevant features. A cascading classifier is used to build the classifier model that classifies the data samples in benign and malignant cases.

3 Background

In this section all techniques and algorithms that are used in this work are explained briefly.

3.1 Synthetic minority over-sampling technique (SMOTE)

SMOTE algorithm is used to resolve class imbalance by synthesizing new observations. Firstly an instance I from minority class is chosen randomly and its nearest neighbors are identified.The synthetic observation is then created by randomly choosing a neighbor N and connecting the instance I and N with a line segment in feature space. Identify a new instance on the line segment convex combination of I and N. The process is repeated for selected points of the minority class.

3.2 Bernoulli process

The Bernoulli process produces a binary output for each possible outcome of a random variable X. There is a sequence of independent Bernoulli trails during this process. For each trail there is probability P that trail results in 1 (success) and probability $1-P$ for this trail to be 0 (failure). The point is that whenever the experiment is performed, it might result in success or failure.

3.3 Artificial bee colony algorithm (ABC)

ABC algorithm presented by Karaboga and is used to solve optimization problems [15, 16]. The algorithm constitutes some basic elements named food sources, employed bees, onlooker bees and scouts bees. Pseudo code of ABC as mentioned in [17] is explained below.

Initialization: Random initialization of food sources is performed as shown in Eq. 1. Each food source reflects a possible solution.

$$x_{ij} = x_{j\min} + \text{rand}(0,1)(x_{j\max} - x_{j\min}) \tag{1}$$

wehere $i = 1,2,3,4,5,6,7,\ldots,N$ and $j = 1,2,3,4,5,6,7,\ldots,D$ such that N is the number of food sources and D is the number of dimensions.

Employed bee: The task of employed bees is to explore the neighbors of the food sources associated with them.The neighbors are generated using Eq. 2.

$$v_{ij} = x_{ij} + \theta_{ij}(x_{ij} - x_{kj}) \tag{2}$$

For each food source x_i a neighbor v_i is generated through the modification of parameter j that is x_{ij} is modified. j and k are random variables where $k = 1,2,3,4,5,6,\ldots$,SN and it must be different from i. Here SN denotes the number of food sources. θ_{ij} is a real number whose value can be -1 to 1. Once the neighbor is generated, its fitness value is calculated by Eq. 3.

$$\text{Fitness}_i = \begin{cases} \dfrac{1}{1+f_i} & \text{if} \quad f_i \geq 0 \\ 1+\text{abs}(f_i) & \text{if} \quad f_i < 0 \end{cases} \tag{3}$$

where f_i is the cost function, if the quality of the neighbor is better than the current food source, then the neighbor becomes the new food source. After all the employed bees have completed their search, the information about the quality of food source is shared with onlooker bees. The probability with which an onlooker bee will choose a food source for further exploration is calculated by Eq. 4.

$$p_i = \frac{\text{Fitness}_i}{\sum_{N=1}^{F} \text{Fitness}_i} \tag{4}$$

Onlooker bees: Bees choose the food sources having a better probability for further exploration of search space. After this, the onlooker bees behave the same as employed bees. Neighbors of chosen food sources are searched, as

explained in the employed bees phase. The food sources having the best fitness is stored.

Scout bees: When a food source is explored the maximum number of times, it is exhausted, and the employed bee changes into a scout bee. The new food source is determined by the scout bee that replaces the exhausted food source.

3.4 CatBoost algorithm

In this algorithm, gradient boosting is performed on decision trees, and they served as base predictors [18–20]. These decision trees use the same splitting criteria across an entire level of the tree. Feature space is partitioned randomly. Each of the feature subsets corresponds to a tree. The splitting criteria used at every level of the tree can be considered pair $p = (k, v)$ with features such as $k = 1, 2, 3, 4, 5 \ldots, m$ with threshold value $v \in R$. By applying the splitting, criteria feature vector X is divided into two disjoint subsets of X^L and X^R for each $x = (x^1, \ldots, x^m) \in X$ we have

$$X = \begin{cases} X^L & \text{if} \quad x^k \leq V \\ X^R & \text{if} \quad x^k \geq V \end{cases} \tag{5}$$

Splitting criteria is defined as

$$\text{argmin}\{P(r, f, M)\} \tag{6}$$

Here M calculates the optimality of both splitting rule r and collection M with regard to target function f, P is given in Eq. 7.

$$P(r, f, M) = \frac{1}{\Sigma_{i=1}^{s} |X_{(i)}|} \left(\Sigma_{i=1}^{s} |X_{(i)}^L| Var\left(f\left(X_{(i)}^L \right) \right) + |X_{(i)}^R| Var\left(f\left(X_{(i)}^R \right) \right) \right) \tag{7}$$

where $f\left(X_{(i)}\right)$ is the score of target samples $X_{(i)}$.

3.5 Multi layer perceptron (MLP)

It is one of the basic algorithms of Deep Learning [21, 22]. In MLP, there is more than one layer of perceptrons. It comprises input, hidden, and output layers. Training the MLP consists of forward or backward pass along with loss function calculation. MLP model predicts the output, and then the loss function is calculated and backpropagated to minimize the error by updating the weights using the gradient.

3.6 Support vector machine (SVM)

SVM is a classification algorithm that separates binary labled training data using hyperplane [23]. If linear separation of the given dataset is not possible, then kernel technique is applied that automatically realizes to nonlinearly map a feature space.

3.7 Random forest (RF)

RF belongs to the class of ensemble learning algorithms [24, 25]. Classification, regression, and other related problems can be solved using Random forest. They combine various decision tree predictors. An assembly of decision trees is produced during the training phase, and the output class is the mean prediction or mode of the classes from each tree. RF is suitable to use with datasets with high dimensionality.

4 Cancer detection through relevant genes identification (CDRGI)

Here we have explained the preprocessing, discrete filtering for feature selection and classification of samples using selected features by cascading classifier.

4.1 Preprocessing

In preprocessing, data are normalized using a trimmed mean of M-values (TMM) [26]. TMM is a useful technique for the normalization of RNA seq data. Synthetic Minority oversampling technique [27] is used to address the high-class imbalance in RNA seq data. This helps in eliminating bias toward the majority class.

4.2 Discrete filtering for feature selection

In an effort to select the optimal number of features (genes) that provide the most significant performance gain for the selected classification algorithm, Discrete Filtering is used. The primary premise of this technique is to eliminate irrelevant features in data. The relevance of features is determined by their influence on the accuracy of the classification model, whereas irrelevance is ascribed to features whose removal or inclusion has the slightest impact on accuracy measure. By selecting features, we can obtain better classification results and reduced execution time to obtained the results. Discrete filtering generates a template, which we called as a filter. This filter can be visualized as a binary string with each bit at different indices are annotated as features (genes). The length of the string and the total number of features are the same. Each bit at a particular index indicates the existence or absence of a feature in data. For example a vector with ten features can be specified as [$f1$, $f2$, $f3$, $f4$, $f5$, $f6$, $f7$, $f8$, $f9$, $f10$] while possible discrete filter can be [0, 0, 0, 1, 0, 1, 0, 1, 1, 0]. It

indicates that features $f4$, $f6$, $f8$, and $f9$ are selected for classification while effectively reducing the number of dimensions, as shown in Fig. 1.

In the proposed technique, certain modifications are made in the classical ABC algorithm. The food sources are initialized using the Bernoulli process [28]. For each index, in the discrete filter, a random number within the range of (0, 1) is generated if the resulting value is less than 0.5, then the related index gets 0 otherwise 1. The selected features are then submitted to the classifier, and the fitness value of the feature subset is determined using the following fitness function.

$$\text{Fitness} = C_{\text{A}} \times \alpha + (1 - \alpha) \times \frac{E_{\text{G}}}{T_{\text{F}}} \tag{8}$$

where parameter α determines the trade-off between the classifier accuracy C_{A} and the number of genes eliminated E_{G} with respect to the number of features T_{F}. Neighbors of the chosen food sources are created by using the Modification Rate (MR) parameter [17]. Neighbors are formed using the Discrete Filter of a current food source. In the classical version of the ABC neighborhood of a food source is determined by perturbation in a single optimization parameter resulting in slower convergence, as shown in Eq. 2. However, in the proposed methodology, discrete filters correspond to optimization parameters, and MR performs their perturbation. For each index of the discrete filter, a random number between 0 and 1 is produced. If this random number is smaller than the perturbation parameter MR, the feature is selected and becomes a member of the subset. In other words, the value at that particular index of the discrete filter is set to 1. Otherwise, the value is not modified. This is represented in equation as follows:

$$X_i = \begin{cases} 1 & \text{if} \quad R_i < \text{MR} \\ X_i & \text{if} \quad \text{Otherwise} \end{cases} \tag{9}$$

where x_i is the index i in a discrete filter. The fitness value for each neighboring feature subset is determined using Eq. 8 to keep the diversity and search the solution space deeply. The crossover operator, as used in the Genetic Algorithm, is applied. Crossover is applied between each current solution and neighboring solution. In this way, two new offsprings are created. After the exploration of employed bees, the solution with the highest fitness value is selected. This solution is then passed to the onlooker bees phase. Once the onlooker bees choose the food source with the best fitness value, it executes the employed bee phase steps. At the end of the onlooker bees phase, the food source with the highest fitness value is stored. These steps are iterated until the termination criteria are reached.

Algorithm 1 Discrete Filtering Pseudocode

SN:Food Source Positions

1: Initialize food sources using Bernoulli process.
2: Submit feature subset (selected food sources) to the classifier and calculate fitness using eq 10.
3: **while** termination criteria is not satisfied **do**
4: **for** $i \leftarrow 1 to SN$ **do**
5: Produce new food sources in neighborhood using MR rate as shown in equation. 8.
6: Perform crossover operation between current solution and neighbouring solution to further explore the search space.
7: Calculate fitness values of all newly generated solutions using equation. 9.
8: Rank the solutions according to fitness values
9: Select the fittest solution
10: **end for**
11: Repeat step 4 to 10 for selected solution
12: **end while**
13: Memorize the best solution
14: Replace the abandoned food source by creating new food source using Bernoulli process

4.3 Cascading classifier

After the feature selection step, the chosen feature subset is passed to the cascading classifier. We used two-stage cascading classifiers, and the first stage CatBoost algorithm is used. All those samples that are classified as noncancerous are passed on to the second stage, where majority voting is done between SVM, RF, and Multi-Layer perceptron classifiers. A sample is termed as cancerous if two of the classifiers at this stage have voted it to be cancerous. Two staged classifiers are built to dual verify the sample class.

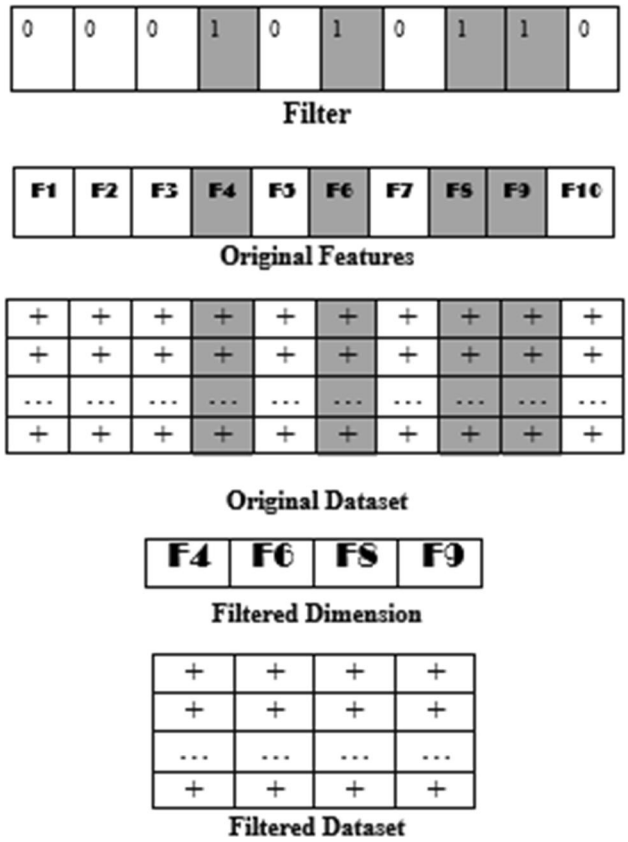

Fig. 1 Feature selection using discrete filtering

5 Experimental results

The Cancer identification framework has been developed using python, and the preprocessing phase is performed using R. Expermients are performed on Inter core i 7 processor with 8 GB memory. Once we obtained the normalized read count for each gene, all the low count genes whose expression levels are very low in all the samples are removed. The proposed approach is trained and evaluated on publicly available datasets, as mentioned in Table 1.

5.1 Evaluation metrics

On account of performance evaluation of presented architecture the evaluation metrics used are discussed in this section. The evaluation metrics, namely average fitness value, Accuracy, Precision, Sensitivity and Specificity used in this work are shown in Eqs. 10, 11, 12, 13 and 14, respectively.

$$\text{Average fitness} = \frac{\text{Fitness}}{\text{Total Number of iterations}} \tag{10}$$

Average fitness values are used to estimate the adequacy of selected features subset with respect to classification of samples using discrete filtering. Average is obtained after iterating five times.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total Number of samples}} \tag{11}$$

Accuracy is the ratio between total correct predictions and number of predictions made in total by classification model.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{12}$$

Precision is the ratio between true positive results in other words that are actually true and total number of positive predictions made either they are true or false.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{13}$$

Sensitivity is estimation of true positive rate. It is calculated by dividing true positive results with all positive data samples.

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \tag{14}$$

Specificity is the true negative rate and determined by dividing true negative data points that are negative in actual with total number of negative data points.

5.2 Discrete filtering

Discrete filtering utilized Binary Artificial Bee Colony (BABC) that is stochastic along with SVM. As mentioned earlier, discrete filters are created to select a subset of features from data. Filters are created heuristically, and their relative fitness is calculated as shown in Table 2. While performing normalization during the Pre-processing phase, we have tested the threshold values on (5, 10, 15) for TMM and found 10, the most suitable threshold value, where we have reduced feature set along with competitive classifier accuracy. The gene expression with a value of

Table 1 Dimensions of datasets used

Dataset	BRCA	STAD	COAD	LIHC	LUSC	KIRC	SKCM
Dimensions	1217 × 60,483	407 × 53,852	512 × 18,984	424 × 60,483	550 × 38,842	607 × 60,483	472 × 60,483

Table 2 Fitness values of selected features

Dataset	Samples	Features passed to BABC-SVM	Average features after shrinkage	Average fitness value of selected features
BRCA	1217	23103	5776	85.197
STAD	407	20538	7251	88.982
COAD	512	20538	5750	88.218
LIHC	424	20125	5635	87.941
LUSC	550	23396	6082	86.521
KIRC	607	22923	6418	82.937
SKCM	472	21999	6160	81.658

Table 3 Results of cascading classifier

Dataset	Accuracy	Precision	Sensitivity	Specificity
BRCA	92.220	86.267	99.739	84.842
STAD	97.412	92.621	99.721	90.021
COAD	97.351	92.831	99.018	90.346
LIHC	90.761	85.021	98.751	82.634
LUSC	98.501	93.603	99.822	91.352
KIRC	95.869	90.314	98.811	90.144
SKCM	89.901	90.544	92.361	92.291

less than 10 is known as pseudogenes (non-coding RNAs) and is considered the least important. The reduced feature set is passed to BABC-SVM based discrete filtering technique.

To determine the effectiveness of selected features cascading classifier is used to classify cancerous samples from normal samples. The performance of the proposed classifier is calculated using above-mentioned metrics.The experiments are run five times and average values are computed and shown in Tables 2 and 3.

5.3 Comparison with deep learning base based tumor type classifier

CDRGI results are compared with [29], as shown in Table 3. It is clear that we have used more samples for each cancer type except for STAD and obtained comparable results. CDRGI outperformed by showing a more significant percentage reduction in features. The average percentage reduction in features by our technique is 72%, and that of reference is 49%, but in terms of accuracy, CDRGI performs better for STAD, COAD, LUSC, and KIRC cancer type detection. For the BRCA, LIHC and SKCM cancer types, counter part algorithm [29] performs better than CDRGI as shown in Table 4.

6 Conclusion

In CDRGI, we have proposed a novel approach to identifying potential cases for each cancer type. The significance of genes in appropriate classification was determined. Most relevant genes were selected using the discrete filtering method. Based on selected genes, the samples were classified using a cascading classifier. Results are compared with seven benchmark datasets. The proposed approach significantly improves the feature sets while performing better in terms of classification accuracy for four cancer types. The CDRGI is useful for large-scale datasets and can

Table 4 Comparison of classification accuracy

Dataset	Samples in CDRGI	Samples in Ref. [29]	Accuracy of CDRGI	Accuracy of Ref. [29]
BRCA	1217	1093	92.22	**99.00**
STAD	407	415	**97.41**	96.00
COAD	512	457	**97.35**	95.00
LIHC	424	371	90.76	**97.00**
LUSC	550	501	**98.50**	91.00
KIRC	607	533	**95.86**	95.00
SKCM	472	469	89.90	**98.00**

Bold highlights the cancer types where CDRGI performed better than counterpart technique

be applied on large dimensions as well. In future work, we will try to cover more cancer types for feature reduction and classification. Also, we will try to add more classifier to check the accuracy of the proposed technique.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Xiao Y, Wu J, Lin Z, Zhao X (2018) A deep learning-based multi-model ensemble method for cancer prediction. Comput Methods Progr Biomed 153:1–9
2. Xiao Y, Wu J, Lin Z, Zhao X (2018) A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data. Comput Methods Progr Biomed 166:99–105
3. Elyasigomari V, Lee DA, Screen HR, Shaheed MH (2017) Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification. J Biomed Inform 67:11–20
4. Khalifa NEM, Taha MHN, Ali DE, Slowik A, Hassanien AE (2020) Artificial intelligence technique for gene expression by tumor RNA-Seq data: a novel optimized deep learning approach. IEEE Access 8:22874–22883
5. Lu H, Chen J, Yan K, Jin Q, Xue Y, Gao Z (2017) A hybrid feature selection algorithm for gene expression data classification. Neurocomputing 256:56–62
6. Cancer-World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cancer
7. Prager GW, Braga S, Bystricky B, Qvortrup C, Criscitiello C, Esin E, Strijbos M (2018) Global cancer control: responding to the growing burden, rising costs and inequalities in access. ESMO Open 3(2):e000285
8. National Cancer Institute. https://www.cancer.gov/about-cancer/understanding/what-is-cancer
9. Saini H, Lal SP, Naidu VV, Pickering VW, Singh G, Tsunoda T, Sharma A (2016) Gene masking-a technique to improve accuracy for cancer classification with high dimensionality in microarray data. BMC Med Genom 9(3):74
10. National Cancer Institute. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
11. Hsu YH, Si D (2018) Cancer Type Prediction and Classification Based on RNA-sequencing Data. In: 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 5374–5377
12. Danaee P, Ghaeini R, Hendrix DA (2017) A deep learning approach for cancer detection and relevant gene identification. In: Pacific symposium on biocomputing 2017. pp 21–229
13. Kashan MH, Nahavandi N, Kashan AH (2012) DisABC: a new artificial bee colony algorithm for binary optimization. Appl Soft Comput 12(1):342–352
14. Lyu B, Haque A (2018) Deep learning based tumor type classification using gene expression data. In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. pp 89–96
15. Hamzeh O, Alkhateeb A, Zheng J, Kandalam S, Rueda L (2020) Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data. BMC Bioinform 21(2):1–10
16. Shon HS, Batbaatar E, Kim KO, Cha EJ, Kim KA (2020) Classification of kidney cancer data using cost-sensitive hybrid deep learning approach. Symmetry 12(1):154
17. Karaboga D (2005) An idea based on honey bee swarm for numerical optimization, vol 200. Technical report-tr06. Erciyes University, Engineering Faculty, Computer Engineering Department, pp 1–10
18. Akay B, Karaboga D (2012) A modified artificial bee colony algorithm for real-parameter optimization. Inf Sci 192:120–142
19. Schiezaro M, Pedrini H (2013) Data feature selection based on artificial bee colony algorithm. EURASIP J Image Video Process 2013(1):47
20. CatBoost. https://catboost.ai/
21. Kang P, Lin Z, Teng S, Zhang G, Guo L, Zhang W (2019) Catboost-based framework with additional user information for social media popularity prediction. In: Proceedings of the 27th ACM international conference on multimedia. pp 2677–2681
22. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2018) CatBoost: unbiased boosting with categorical features. In: Advances in neural information processing systems. pp 6638–6648
23. Understanding of MultiLayer (MLP) Perceptron. https://medium.com/@AI.with.Kain/understanding-of-multilayer-perceptron-mlp-8f179c4a135f
24. Tang J, Deng C, Huang GB (2015) Extreme learning machine for multilayer perceptron. IEEE Trans Neural Netw Learn Syst 27(4):809–821
25. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10):906–914
26. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
27. https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/
28. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11(3):R25
29. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357