Article

Dimension Reduction and Classifier-Based Feature Selection for Oversampled Gene Expression Data and Cancer Classification

Olutomilayo Olayemi Petinrin [1] **, Faisal Saeed** [2,*] **, Naomie Salim** [3] **, Muhammad Toseef** [1] **, Zhe Liu** [1]
and Ibukun Omotayo Muyide [4]

[1] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong;
opetinrin2-c@my.cityu.edu.hk (O.O.P.); mtoseef2-c@my.cityu.edu.hk (M.T.); zliu39-c@my.cityu.edu.hk (Z.L.)

[2] DAAI Research Group, Department of Computing and Data Science,
School of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, UK

[3] UTM Big Data Centre, Ibnu Sina Institute for Scientific and Industrial Research,
Universiti Teknologi Malaysia, Johor Bahru 81310, Johor, Malaysia; naomie@utm.my

[4] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA; imuyide3@gatech.edu

[*] Correspondence: faisal.saeed@bcu.ac.uk

Abstract: Gene expression data are usually known for having a large number of features. Usually, some of these features are irrelevant and redundant. However, in some cases, all features, despite being numerous, show high importance and contribute to the data analysis. In a similar fashion, gene expression data sometimes have limited instances with a high rate of imbalance among the classes. This can limit the exposure of a classification model to instances of different categories, thereby influencing the performance of the model. In this study, we proposed a cancer detection approach that utilized data preprocessing techniques such as oversampling, feature selection, and classification models. The study used SVMSMOTE for the oversampling of the six examined datasets. Further, we examined different techniques for feature selection using dimension reduction methods and classifier-based feature ranking and selection. We trained six machine learning algorithms, using repeated 5-fold cross-validation on different microarray datasets. The performance of the algorithms differed based on the data and feature reduction technique used.

Keywords: cancer classification; gene expression; machine learning; microarray data; sampling methods

1. Introduction

Gene expression involves the translation of information encoded in a gene into gene products, including proteins, tRNA, rRNA, mRNA, or snRNA. With the increase in technological standards, gene expression continues to have a sporadic increase in importance for health and life science applications [1]. Prognostic risk scores from gene expression have shown prominent clinical values as they are promising biomarkers. They can be used for the prediction of prognosis, including the identification of mortality and metastasis risks in patients. They can also be used to determine the response of patients to treatment [2]. Identifying the risk of cancer recurrence or metastasis in patients can help clinicians strategically recommend effective treatment. Furthermore, the determination of response to treatment can identify the overall survival of the patients and intuitively develop novel drugs or appropriate treatment based on each patient's classification. In the majority of cancer types, HLA gene expression has been shown to prolong overall survival [3]. On the other hand, an increase in the expression of Human Endogenous Retrovirus K mRNA in the blood is linked to the presence of breast cancer, which shows it is a biomarker [4]. BRCA2 is another gene whose expression is associated with highly proliferative and aggressive breast cancer. The higher the expression of BRCA2, the more aggressiveness the breast cancer [5]. This indicates its potential as a biomarker for breast cancer. In essence,

biological determinants such as predictive gene expression signatures can now be used for the effective classification of tumors according to their subgroup [6]. The profiling of gene expression for breast cancer or other cancer types can be further improved using clinicopathological and microenvironmental features [7–9].

The generation of data in the biomedical and health fields has increased sporadically while yielding samples with a high number of features [10,11]. The challenge of high-dimensional data is the difficulty associated with manual analysis and the redundancy that comes along with some of the features. Over the years, several studies have been carried out based on feature selection. In [12], the authors utilized a hybrid filter and wrapper method for the selection of features in gene expression data. The authors also used LASSO, an embedded technique, and reported that the performance of the machine learning algorithms was better with the implementation of LASSO on the examined high dimensional datasets. Townes et al. [13] implemented a simple multinomial method using generalized principal component analysis and carried out feature selection using deviance. Different combinations of methods were further compared with current methods to show their performance. Evolutionary algorithms have also been implemented in some studies for the improvement of feature selection. Jain et al. [14] implemented a correlation-based feature selection method improved with binary particle swarm optimization for the selection of genes before classifying the cancer types using Naïve Bayes algorithm. This method improved the classifier's performance. In the same vein, Kabir et al. [15] compared two different dimension reduction techniques—PCA, and autoencoders for the selection of features in a prostate cancer classification analysis. Two machine learning methods—neural networks and SVM—were further used for classification. The study showed that the classifiers performed better on the reduced dataset.

Prasad et al. [16] used a recursive particle swarm optimization technique with the integration of filter-based methods for ranking. The authors also reported improved performance based on five datasets. Another similar approach for gene selection involves the hybridization of ant colony optimization and cellular learning automata. Based on the ROC curve evaluation of three classifiers, the proposed method selected the minimum gene needed for maximum performance of the classifiers [17]. Similarly, Alhenawi et al. [18] proposed a hybrid feature selection technique using Hill Climbing, the Novel LS algorithm, and Tabu search for microarray data. This is similar to the filter-wrapper and embedded technique utilized for gene expression data in [12]. However, Keshta et al. [19] proposed a multi-stage algorithm for the extraction and selection of features in a cancer detection study. It was reported that despite the reduction in the number of features used for classification, the performance of classifiers was either enhanced or unchanged. In addition, a nested genetic algorithm consisting of an outer genetic algorithm and an inner genetic algorithm has previously been implemented for the gene selection of a colon and lung dataset using 5-fold cross-validation [20]. A significant increase in classification accuracy was also reported. Several other feature/gene selection techniques are being improved and implemented to improve the accuracy of cancer classification. These include robust linear discriminant analysis [21], adaptive principal component analysis [22], and the use of deep variational autoencoders, especially in studies that involve the use of deep learning [23].

In this study, we considered the problem of imbalanced data, which is common in health data, before using dimension reduction techniques such as principal component analysis (PCA), truncated singular value decomposition (TSVD), and T-stochastic neighbor embedding (TSNE) to address the high dimensionality issue peculiar to gene expression data. We also utilized the ability of some machine learning algorithms to rank some genes and make selections based on a specified threshold.

2. Materials and Methods

2.1. Dataset Description

The six gene expression datasets used in this study are the brain, colon, leukemia, lymphoma, prostate, and small blue round cell tumor (SBRCT) datasets, as explained in

our previous work [24] and shown in Table 1. Forty-two (42) patient samples make up the brain cancer microarray dataset. The tumors include 10 medulloblastomas, 5 atypical tera-toid/rhabdoid tumors of the central nervous system (CNS), 5 rhabdoid tumors of the renal and extrarenal organs, 8 supratentorial primitive neuroectodermal tumors (PNETs), 10 non-embryonal brain tumors, and 4 normal human cerebella. 6817 genes are present in the first oligonucleotide microarrays. They underwent thresholding during pre-processing by [25]. Therefore, for the entire dataset with five distinct sample classes, there are 5597 genes. Alon et al. [26] conducted the initial analysis of the colon cancer microarray dataset. The raw data from the Affymetrix oligonucleotide arrays was processed by the dataset's original authors. Samples of both normal and tumor tissue make up the dataset. The total number of samples is 62, and the 2000 gene numbers after pre-processing reported by earlier authors [27,28] are the total gene numbers. Acute lymphoblastic leukemia and acute myeloid leukemia are the two kinds of acute leukemia studied for gene expression, whose results were used to create the leukemia cancer dataset. Affymetrix high-density oligonucleotide arrays, which had 6817 genes but were reduced to 3051 genes and further analyzed by [29], were used to determine the levels of gene expression. 47 cases of ALL (38 B-cell ALL and 9 T-cell ALL) and 25 instances of AML make up the dataset. Dudoit et al. [30] performed more pre-processing on the dataset. The dataset can be acquired from [27,28].

Table 1. Detailed Information about the datasets.

Dataset	Classes	Instances	Attributes
Brain	5	42	5598
Colon	2	62	2001
Leukemia	2	72	3572
Lymphoma	3	62	4027
Prostate	2	102	6034
SBRCT	4	63	2309

The dataset for the lymphoma microarray is found in [25]. It comprises 62 samples and 4026 genes. The majority of the data samples are from three distinct adult lymphoid malignancies: 42 samples represent diffuse large B-cell lymphoma (DLBCL), 9 samples come from follicular lymphoma (FL), and 11 samples come from chronic lymphocytic leukemia (CLL). The dataset is also available can be found in the literature [27,28]. 50 of the samples in the prostate cancer dataset are normal prostate specimens, while the remaining 52 are tumors. The collection contains 102 patterns of gene expression. About 12,600 genes make up this microarray dataset, which is based on an oligonucleotide microarray. The dataset still contains 6033 genes after pre-processing [31]. There are four distinct classifications in the Small Round Blue-Cell Tumor (SRBCT) microarray dataset, which initially had 6567 genes and 63 samples. Whereas, 8 samples come from NHL, 12 from NB, 20 from the RMS, and 23 from EWS. The number of genes decreased to 2308 after pre-processing. This dataset was produced using [32] and is available in [27,28].

2.2. Methods

From the information about the datasets contained in Table 1, there is a clear indication of the high dimensionality of the datasets. This high dimensionality informs the basics of the feature selection and dimension reduction methods used in the study. Before analysis, we normalized the data using the min-max normalization method. The formula for min-max normalization is given as $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$, where x is the vector in a feature column, x_{min} is the minimum value in the column x_{max} and is the maximum value in the column. Because the analysis deals with the selection of features, it is crucial to balance the features so that a feature does not have more contributing capacity to the analysis than another column. For further analysis, we note the imbalance of the data.

For health-related data, it is always essential to deal with the imbalance in order to present a false representation of the evaluation, especially using accuracy. Subsequently,

oversampling was carried out on each of the datasets using the SVMSMOTE technique [24]. This technique uses the support vector machine to predict new and unknown samples around the borderline. In a support vector machine, the borderline or margin is crucial for the establishment of the decision boundary. This is why SVMSMOTE focuses on instances of the minority class that are found along the borderline and therefore generates more samples in such a way that new instances of the minority class will appear where there are fewer instances of the majority class. In this study, the nearest neighbor parameter was set to 3 and the random state to 42.

Two categories of feature reduction methods were considered in this study. The first one entails the use of dimension reduction methods, and the second category entails the use of classifiers for feature ranking and selection. The dimension reduction methods used are principal component analysis (PCA), truncated singular value decomposition (TSVD), and T-distributed stochastic neighbor embedding (TSNE). On the other hand, the classifiers used for ranking are random forest (RF) and logistic regression (LR). PCA tries to maintain the data variance as much as possible. Basis vectors, also known as principal components, along with the maximum variance of the data are chosen with the goal of minimizing the reconstruction error over all the data points. To do this, the eigenvectors with the largest eigenvalues are selected first. TSVD uses a matrix factorization technique that is similar to PCA. However, TSVD is performed on the data matrix, while PCA is performed on the covariance matrix. The name "truncated" comes from the number of columns being equal to the truncation. That is, matrices with a specified number of columns are produced. The primary goal of TSNE is to preserve pairwise similarities as much as possible. However, like PCA, it does not maintain inputs. This makes it useful for visualization and exploration. It treats similarities in the original space as probabilities and finds the embedding that preserves probability structure. TSNE uses Kullback-Leibler (KL) divergence as a measure of the similarity between two probability distributions.

All examined models went through a 5-fold cross-validation repeated for a total of five replicates and were evaluated based on four different metrics: accuracy, precision, recall, and F1-score. These evaluation criteria are calculated based on Equations (1)–(4) respectively.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F1-score} = \frac{2*\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP, FP, TN, and FN are true positive, false positive, true negative, and false negative, respectively.

We implemented all analyses in this study on a Windows 10 64-bit operating system on a x64-based processor computer with 64 GB of RAM. The processor specification is an Intel (R) Core (TM) i7-9700 K CPU @ 3.6 GHz. Python 3.6 and libraries, including scikit-learn, were utilized for the analyses. Figure 1 further shows the procedure for study analysis.
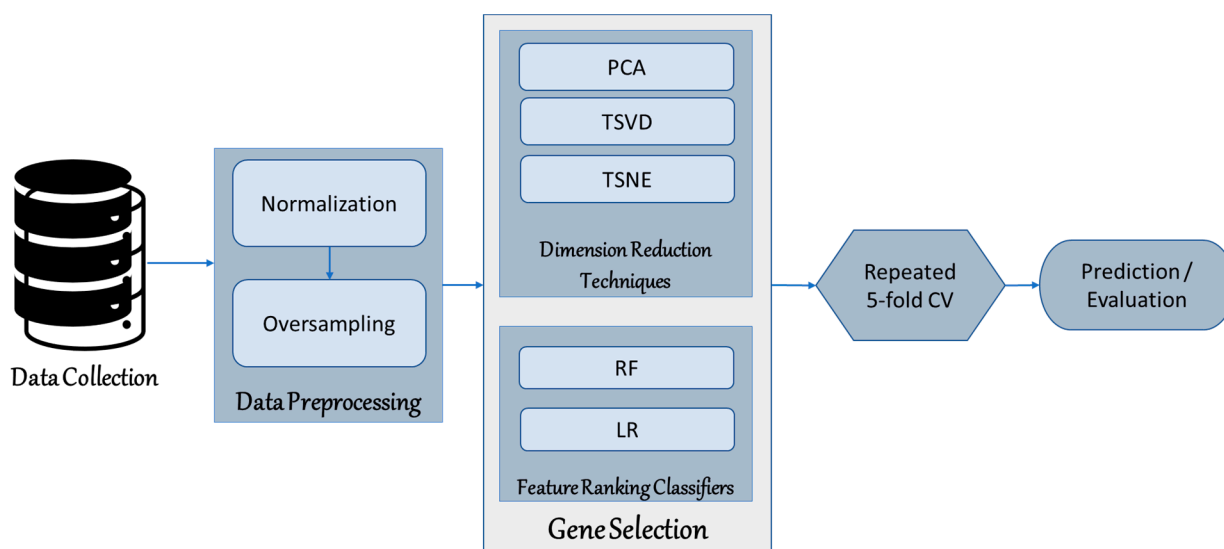
Figure 1. Process Diagram of the Study.

3. Results and Discussions

For all analyses, we used repeated 5-fold cross-validation. The 5-fold cross-validation had the parameter shuffle set to true and was repeated five times to generate 25 results per evaluation metric. The mean of each analysis is reported as the result, while a full report of the result with the standard deviation is given in the Supplementary Material. For this study, we examined the different feature reduction techniques using six different classifiers, logistic regression (LR), random forest (RF), support vector machine (SVM), gradient boosting classifier (GBC), Gaussian Naïve Bayes (GNB), and k-nearest neighbor (KNN). These are all commonly used classifiers that behave differently, which has thus influenced their choice for the analyses. For the logistic regression, the parameter for maximum iteration was set to 500, and liblinear was selected as the solver. In a similar manner, for the random forest and gradient boosting classifiers, the number of estimators was 500. For the support vector machine, Gaussian naïve bayes, and k-nearest neighbor, the default sklearn parameters were used.

3.1. Performance of Classification Methods after Oversampling

Firstly, we carried out the analysis on all the datasets without any prior oversampling or reduction. The only preprocessing technique that was applied was normalization. We further compared the result with the result generated after the SVMSMOTE oversampling technique was used. The results from the two analyses are shown in Tables 2–5. The better result is highlighted. From the tables, we deduce that the majority of the analysis with oversampling has better performance than the one without oversampling. A peculiar benefit of oversampling is that it allows the model to be exposed and trained with a balanced number of both the majority and minority samples. For the lymphoma dataset, there is consistency between the original dataset and the oversampled dataset using random forest and support vector machines.

Table 2. Comparison between the accuracy of models before and after oversampling.

Method	SVMSMOTE STATUS	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
LR		0.8167	0.8872	0.9857	**1.0000**	0.9110	0.9846
	With	**0.8667**	**0.9000**	**0.9895**	0.9909	**0.9229**	**0.9889**
RF		0.7444	0.8410	0.9457	**1.0000**	0.8910	0.9846
	With	**0.8444**	**0.9250**	**0.9895**	**1.0000**	**0.8943**	**0.9889**

Table 2. *Cont.*

Method	SVMSMOTE STATUS	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
SVM		0.6750	0.8564	0.9857	**1.0000**	0.8714	0.8897
	With	**0.7600**	**0.9125**	**1.0000**	**1.0000**	**0.9038**	**0.9234**
GBC		**0.6383**	0.7646	0.8486	**0.9833**	**0.8818**	0.8297
	With	0.6111	**0.8000**	**0.8957**	0.9814	0.8762	**0.8457**
GNB		**0.6417**	0.8551	0.9724	0.9167	**0.6371**	0.9205
	With	0.6311	**0.8625**	**1.0000**	**0.9909**	0.6057	**0.9778**
KNN		0.6972	0.7923	**0.9305**	0.9846	**0.8614**	0.7744
	With	**0.7800**	**0.8750**	0.8830	0.9727	0.8552	**0.8146**

Table 3. Comparison between the precision of models before and after oversampling.

Method	SVMSMOTE STATUS	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
LR		0.7783	0.8737	0.9889	**1.0000**	0.9178	**0.9929**
	With	**0.8567**	**0.9056**	**0.9909**	0.9905	**0.9324**	**0.9929**
RF		0.6433	0.8304	0.9556	**1.0000**	**0.9053**	**0.9929**
	With	**0.8567**	**0.9362**	**0.9909**	**1.0000**	0.9013	**0.9929**
SVM		0.5483	0.8438	0.9889	**1.0000**	0.8757	0.8625
	With	**0.8033**	**0.9249**	**1.0000**	**1.0000**	**0.9086**	**0.9200**
GBC		0.5615	0.7007	0.8378	**0.9905**	**0.8920**	0.7827
	With	**0.6340**	**0.8144**	**0.8956**	0.9771	0.8836	**0.8552**
GNB		**0.6225**	0.8421	0.9746	0.9016	**0.6630**	0.9125
	With	0.6080	**0.8706**	**1.0000**	**0.9917**	0.6305	**0.9762**
KNN		0.5258	0.7814	**0.9139**	**0.9833**	**0.8669**	**0.8177**
	With	**0.6367**	**0.9072**	0.9043	0.9742	0.8545	0.7935

Table 4. Comparison between the recall of models before and after oversampling.

Method	SVMSMOTE STATUS	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
LR		0.7850	0.8873	0.9833	**1.0000**	0.9048	0.9750
	With	**0.8750**	**0.8978**	**0.9889**	0.9944	**0.9149**	**0.9833**
RF		0.7150	0.8339	0.9500	**1.0000**	**0.8923**	**0.9875**
	With	**0.8617**	**0.9246**	**0.9889**	**1.0000**	0.8918	0.9833
SVM		0.6567	0.8506	0.9833	**1.0000**	0.8690	0.8564
	With	**0.8017**	**0.9103**	**1.0000**	**1.0000**	**0.8995**	**0.9324**
GBC		0.5473	0.7208	0.8739	0.9778	**0.8784**	0.7939
	With	**0.6283**	**0.8016**	**0.8958**	**0.9849**	0.8733	**0.8501**
GNB		0.6283	0.8530	0.9722	0.8533	**0.6405**	0.8806
	With	**0.7167**	**0.8585**	**1.0000**	**0.9926**	0.6167	**0.9733**
KNN		0.6400	0.7796	**0.9444**	**0.9926**	**0.8606**	0.8025
	With	**0.7567**	**0.8714**	0.8888	0.9778	0.8474	**0.8097**

Table 5. Comparison between F1 scores of models before and after oversampling.

Method	SVMSMOTE STATUS	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
LR		0.7480	0.8736	0.9850	**1.0000**	0.9079	0.9795
	With	**0.8427**	**0.8978**	**0.9894**	0.9920	**0.9185**	**0.9862**
RF		0.6467	0.8228	0.9450	**1.0000**	0.8895	**0.9890**
	With	**0.8307**	**0.9232**	**0.9894**	**1.0000**	**0.8904**	0.9862
SVM		0.5610	0.8405	0.9850	**1.0000**	0.8695	0.8448
	With	**0.7573**	**0.9097**	**1.0000**	**1.0000**	**0.8996**	**0.9149**
GBC		0.5288	0.6944	0.8205	**0.9815**	**0.8776**	0.7674
	With	**0.5730**	**0.7979**	**0.8898**	0.9794	0.8715	**0.8394**
GNB		0.5793	0.8408	0.9715	0.8608	**0.6310**	0.8840
	With	**0.6260**	**0.8598**	**1.0000**	**0.9916**	0.5958	**0.9706**
KNN		0.5467	0.7700	**0.9158**	**0.9866**	**0.8594**	0.7650
	With	**0.6660**	**0.8665**	0.8759	0.9744	0.8497	**0.7711**

3.2. Performance of Classification Methods Based on Dimension Reduction Techniques

Due to the better performance of the models using the oversampled data in Tables 2–5, the same data was used for the rest of the analyses. In Tables 6–9, we compared the performance of the trained models based on three-dimensional reduction methods, namely principal component analysis (PCA), truncated singular value decomposition (TSVD), and t-distributed stochastic neighbor embedding (TSNE). The parameters of PCA and TSVD were set to make the cumulative explained variance of the chosen principal components 0.99. For TSNE, the number of components was set at 3, and perplexity was set at 50. From the results shown in Tables 6–9, the performance of PCA and TSVD is relatively similar, although in many cases, PCA had better performance. In the majority of the analyses, the performance of the PCA or TSVD dimension-reduced analyses was better than the performance of the classifiers before reduction. TSNE, on the other hand, had generally poor performance for all the datasets and classifiers examined. We suppose that this can be attributed to the fact that TSNE has random probability and attempts to retain the variance of neighboring points, that is, local variance. In many cases, TSNE is used just for data visualization, especially for 2D or 3D visualization of images while retaining the local variance.

Table 6. Comparison between model accuracy for PCA, TSVD, and TSNE using an oversampled dataset.

Method	FS Method	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
LR	PCA	**0.9111**	**0.9000**	**0.9895**	**0.9909**	**0.9229**	**0.9889**
	TSVD	0.8667	**0.9000**	**0.9895**	**0.9909**	**0.9229**	**0.9889**
	TSNE	0.2378	0.5500	0.6497	0.6139	0.4990	0.2731
RF	PCA	**0.7822**	0.8250	**0.8842**	0.9545	0.8176	**0.8591**
	TSVD	0.7156	**0.8750**	**0.8842**	**1.0000**	**0.8271**	0.8164
	TSNE	0.2378	0.6500	0.6696	0.6229	0.4605	0.4041
SVM	PCA	**0.7356**	**0.9250**	**1.0000**	**1.0000**	**0.9133**	0.9012
	TSVD	0.5067	0.9000	**1.0000**	0.9909	0.8457	**0.9123**
	TSNE	0.2178	0.6000	0.6918	0.6961	0.4795	0.4047

Table 6. *Cont.*

Method	FS Method	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
GBC	PCA	0.6422	0.7850	**0.8942**	**0.9537**	0.8210	**0.8668**
	TSVD	**0.6644**	**0.8125**	0.8731	**0.9537**	**0.8290**	0.8351
	TSNE	0.2800	0.6125	0.6744	0.4667	0.4463	0.3625
GNB	PCA	**0.7578**	0.8250	0.8415	0.9087	**0.6533**	**0.7281**
	TSVD	0.7356	**0.8750**	**0.8520**	**0.9268**	0.6433	0.6953
	TSNE	0.2356	0.6125	0.6386	0.6043	0.4405	0.3497
KNN	PCA	**0.7822**	**0.8750**	**0.9041**	**0.9727**	0.8452	0.8041
	TSVD	0.7800	**0.8750**	0.8830	**0.9727**	**0.8652**	**0.8146**
	TSNE	0.2600	0.5875	0.6497	0.6697	0.4429	0.3708

Table 7. Comparison between model precision for PCA, TSVD, and TSNE using an oversampled dataset.

Method	FS Method	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
LR	PCA	**0.8900**	**0.9056**	**0.9909**	**0.9905**	**0.9324**	**0.9929**
	TSVD	0.8567	**0.9056**	**0.9909**	**0.9905**	**0.9324**	**0.9929**
	TSNE	0.1660	0.5568	0.6548	0.6207	0.5084	0.2679
RF	PCA	**0.8250**	0.8460	**0.9054**	0.9744	0.8187	**0.8771**
	TSVD	0.7283	**0.8888**	**0.9054**	**1.0000**	**0.8234**	0.8290
	TSNE	0.2233	0.6658	0.6603	0.6270	0.4732	0.4012
SVM	PCA	**0.7817**	**0.9434**	**1.0000**	**1.0000**	**0.9239**	0.8925
	TSVD	0.4844	0.9056	**1.0000**	0.9905	0.8468	**0.9067**
	TSNE	0.1647	0.6076	0.6880	0.6946	0.4736	0.4241
GBC	PCA	0.6143	0.8075	**0.8913**	**0.9619**	0.8384	**0.8720**
	TSVD	**0.6407**	**0.8190**	0.8688	0.9528	**0.8563**	0.8367
	TSNE	0.2397	0.6213	0.6882	0.4888	0.4560	0.3752
GNB	PCA	**0.7590**	0.8417	0.8626	0.9331	**0.6732**	**0.7521**
	TSVD	0.7123	**0.8999**	**0.8756**	**0.9466**	0.6646	0.7282
	TSNE	0.1800	0.6484	0.6397	0.6034	0.4400	0.3502
KNN	PCA	**0.6767**	**0.9072**	**0.9157**	**0.9742**	0.8449	0.7810
	TSVD	0.6367	**0.9072**	0.9043	**0.9742**	**0.8658**	**0.7935**
	TSNE	0.2200	0.5889	0.6600	0.6743	0.4497	0.3923

Table 8. Comparison between model recall for PCA, TSVD, and TSNE using an oversampled dataset.

Method	FS Method	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
LR	PCA	**0.9083**	**0.8978**	**0.9889**	**0.9944**	**0.9149**	**0.9833**
	TSVD	0.8750	**0.8978**	**0.9889**	**0.9944**	**0.9149**	**0.9833**
	TSNE	0.2133	0.5528	0.6648	0.5949	0.5106	0.2908
RF	PCA	**0.8517**	0.8274	**0.8929**	0.9630	0.8157	**0.8844**
	TSVD	0.7717	**0.8792**	**0.8929**	**1.0000**	**0.8195**	0.8427
	TSNE	0.2617	0.6581	0.6566	0.6036	0.4738	0.4147
SVM	PCA	**0.7817**	**0.9214**	**1.0000**	**1.0000**	**0.9072**	0.9074
	TSVD	0.6100	0.8978	**1.0000**	0.9944	0.8459	**0.9233**
	TSNE	0.2133	0.5986	0.6941	0.7038	0.4873	0.4318
GBC	PCA	0.6413	0.7848	**0.8993**	0.9589	0.8061	**0.8849**
	TSVD	**0.6613**	**0.8140**	0.8771	**0.9626**	**0.8101**	0.8378
	TSNE	0.2440	0.6188	0.6924	0.4256	0.4572	0.3864
GNB	PCA	**0.8017**	0.8242	0.8418	0.8804	**0.6506**	**0.7124**
	TSVD	0.7617	**0.8728**	**0.8518**	**0.9081**	0.6415	0.6736
	TSNE	0.2533	0.6139	0.6479	0.5776	0.4488	0.3746
KNN	PCA	**0.7867**	**0.8714**	**0.9099**	**0.9778**	0.8363	0.8035
	TSVD	0.7567	**0.8714**	0.8888	**0.9778**	**0.8565**	**0.8097**
	TSNE	0.2667	0.5897	0.6628	0.6759	0.4473	0.4254

Table 9. Comparison between model F1 scores for PCA, TSVD, and TSNE using an oversampled dataset.

Method	FS Method	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
LR	PCA	**0.8827**	**0.8978**	**0.9894**	**0.9920**	**0.9185**	**0.9862**
	TSVD	0.8427	**0.8978**	**0.9894**	**0.9920**	**0.9185**	**0.9862**
	TSNE	0.1780	0.5457	0.6417	0.5811	0.4912	0.2504
RF	PCA	**0.7923**	0.8210	**0.8806**	0.9585	0.8131	**0.8533**
	TSVD	0.6977	**0.8740**	**0.8806**	**1.0000**	**0.8206**	0.8076
	TSNE	0.2180	0.6458	0.6488	0.5881	0.4591	0.3809
SVM	PCA	**0.7373**	**0.9219**	**1.0000**	**1.0000**	**0.9092**	0.8871
	TSVD	0.4842	0.8978	**1.0000**	0.9920	0.8422	**0.9042**
	TSNE	0.1615	0.5927	0.6809	0.6750	0.4665	0.4004
GBC	PCA	0.5804	0.7791	**0.8903**	**0.9580**	0.8110	**0.8597**
	TSVD	**0.5988**	**0.8110**	0.8684	0.9542	**0.8152**	0.8278
	TSNE	0.2245	0.6102	0.6666	0.4306	0.4398	0.3361
GNB	PCA	**0.7522**	0.8209	0.8342	0.8866	**0.6387**	**0.7102**
	TSVD	0.7056	**0.8699**	**0.8442**	**0.9125**	0.6274	0.6859
	TSNE	0.1880	0.6026	0.6297	0.5553	0.4318	0.3318
KNN	PCA	**0.7003**	**0.8665**	**0.8992**	**0.9744**	0.8389	0.7601
	TSVD	0.6660	**0.8665**	0.8759	**0.9744**	**0.8594**	**0.7711**
	TSNE	0.2227	0.5751	0.6376	0.6491	0.4394	0.3804

3.3. Performance of Classification Methods Based on Classifier-Based Gene Ranking and Selection

Furthermore, we used two classifiers (random forest and logistic regression) for feature ranking and selection. The oversampled datasets were also used for analyses. For the feature selection based on random forest, we used 500 estimators as the parameters and only selected features that were above the mean threshold. On the other hand, for feature selection based on logistic regression, only features above the median threshold for each were selected. We have employed different thresholds for the two techniques to discover if there would be a significant difference in the performance of the classifiers based on the threshold of the feature selection. For the Lymphoma dataset, both random forest and logistic regression had the same performance across evaluations. This was similarly noticed with the small blue round cell tumor (SBRCT) dataset, except with the use of gradient boosting classifiers, and k-nearest neighbor. Overall, Tables 10–13 show that both classifiers had good performance in the ranking and selection of features, although they utilize different strategies for their threshold.

Table 10. Comparison between model accuracy for RF and LR using an oversampled dataset.

Method	FS Method	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
LR	RF	0.9333	0.9000	0.9895	**1.0000**	0.9514	**0.9889**
	LR	**0.9556**	**0.9125**	**1.0000**	**1.0000**	**0.9610**	**0.9889**
RF	RF	**0.8889**	0.9000	0.9895	**1.0000**	**0.9324**	**0.9889**
	LR	0.8667	**0.9250**	**1.0000**	**1.0000**	0.9229	**0.9889**
SVM	RF	0.9111	0.9125	**1.0000**	**1.0000**	**0.9324**	**0.9889**
	LR	**0.9333**	**0.9375**	**1.0000**	**1.0000**	0.9229	**0.9889**
GBC	RF	**0.6289**	**0.8600**	**0.8978**	**0.9814**	**0.8686**	**0.8591**
	LR	0.5889	0.8025	0.8957	**0.9814**	0.8662	0.8568
GNB	RF	0.7178	0.9000	**1.0000**	**1.0000**	**0.8357**	**0.9889**
	LR	**0.7400**	**0.9250**	**1.0000**	**1.0000**	0.7110	**0.9889**
KNN	RF	0.8667	0.9000	**0.9895**	0.9909	**0.9229**	**0.9784**
	LR	**0.9556**	**0.9125**	0.9778	0.9909	0.8938	0.9673

Table 11. Comparison between model precision for RF and LR using an oversampled dataset.

Method	FS Method	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
LR	RF	0.9433	0.9056	0.9909	**1.0000**	0.9629	**0.9929**
	LR	**0.9633**	**0.9249**	**1.0000**	**1.0000**	**0.9691**	**0.9929**
RF	RF	0.8767	0.9056	0.9909	**1.0000**	**0.9428**	**0.9929**
	LR	**0.8767**	**0.9362**	**1.0000**	**1.0000**	0.9324	**0.9929**
SVM	RF	0.8900	0.9249	**1.0000**	**1.0000**	**0.9428**	**0.9929**
	LR	**0.9500**	**0.9516**	**1.0000**	**1.0000**	0.9324	**0.9929**
GBC	RF	**0.6759**	**0.8754**	**0.8963**	**0.9771**	**0.8770**	**0.8685**
	LR	0.6270	0.8103	0.8956	**0.9771**	0.8758	0.8594
GNB	RF	0.7200	0.9056	**1.0000**	**1.0000**	**0.8385**	**0.9929**
	LR	**0.7667**	**0.9335**	**1.0000**	**1.0000**	0.7219	**0.9929**
KNN	RF	0.8167	0.9056	**0.9909**	0.9905	**0.9344**	**0.9804**
	LR	**0.9633**	**0.9266**	0.9714	0.9905	0.9012	0.9720

Table 12. Comparison between model recall for RF and LR using an oversampled dataset.

Method	FS Method	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
LR	RF	0.9417	0.8978	0.9889	**1.0000**	0.9428	**0.9833**
	LR	**0.9617**	**0.9103**	**1.0000**	**1.0000**	**0.9553**	**0.9833**
RF	RF	0.8883	0.8978	0.9889	**1.0000**	**0.9226**	**0.9833**
	LR	**0.8750**	**0.9246**	**1.0000**	**1.0000**	0.9149	**0.9833**
SVM	RF	0.9083	0.9103	**1.0000**	**1.0000**	**0.9226**	**0.9833**
	LR	**0.9483**	**0.9357**	**1.0000**	**1.0000**	0.9149	**0.9833**
GBC	RF	**0.6657**	**0.8560**	**0.8990**	0.9849	**0.8649**	**0.8611**
	LR	0.6117	0.8041	0.8958	**0.9849**	0.8603	0.8591
GNB	RF	0.7850	0.8978	**1.0000**	**1.0000**	**0.8297**	**0.9833**
	LR	**0.7983**	**0.9228**	**1.0000**	**1.0000**	0.7156	**0.9833**
KNN	RF	0.8633	0.8978	**0.9889**	**0.9944**	**0.9101**	**0.9762**
	LR	**0.9617**	**0.9107**	0.9846	**0.9944**	0.8819	0.9662

Table 13. Comparison between model F1 scores for RF and LR using an oversampled dataset.

Method	FS Method	Brain	Colon	Leukemia	Lymphoma	Prostate	SBRCT
LR	RF	0.9253	0.8978	0.9894	**1.0000**	0.9481	**0.9862**
	LR	**0.9520**	**0.9097**	**1.0000**	**1.0000**	**0.9592**	**0.9862**
RF	RF	**0.8640**	0.8978	0.9894	**1.0000**	**0.9279**	**0.9862**
	LR	0.8493	**0.9232**	**1.0000**	**1.0000**	0.9185	**0.9862**
SVM	RF	0.8827	0.9097	**1.0000**	**1.0000**	**0.9279**	**0.9862**
	LR	**0.9360**	**0.9356**	**1.0000**	**1.0000**	0.9185	**0.9862**
GBC	RF	**0.6066**	**0.8570**	**0.8926**	0.9794	**0.8637**	**0.8515**
	LR	0.5470	0.8014	0.8898	**0.9794**	0.8591	0.8481
GNB	RF	0.7063	0.8978	**1.0000**	**1.0000**	**0.8302**	**0.9862**
	LR	**0.7320**	**0.9237**	**1.0000**	**1.0000**	0.7075	**0.9862**
KNN	RF	0.8153	0.8978	**0.9894**	**0.9920**	**0.9170**	**0.9752**
	LR	**0.9520**	**0.9106**	0.9750	**0.9920**	0.8879	0.9651

4. Conclusions

Data generated in the medical and bioinformatics fields are known to have a large number of features. Analyzing these features manually is exhausting, and this is where the utilization of machine learning tools comes in handy. Another major issue with health data is the high rate of class imbalance. This sometimes affects the integrity and robustness of models built using such data. Models might be unable to accurately predict the class of unseen instances in the minority class. In this study, we have considered SVMSMOTE for oversampling the data and thereby increasing the number of instances of each data point. We discover that, generally, the performance of the examined models was better after oversampling. The application of the dimension reduction techniques and feature ranking and selection further improved the performance in many instances. Principal component analysis and truncated singular value decomposition performed better than t-distributed stochastic neighbor embedding. TSNE in fact had a generally bad performance, as shown in Tables 6–9. It is advised to use TSNE primarily for dimension reduction for data visualization. In the same vein, both random forest and logistic regression classifiers were effective in the selection of features despite utilizing different threshold criteria. Although

a low number of analyses with the original dataset had good performance, the majority of the analyses with SVMSMOTE and feature reduction had better performance, saved time, and enhanced the interpretability of models. Future works will investigate the analysis of gene expression data and cancer classification using different oversampling and dimension reduction methods on different microarray datasets.

References

1. Thakur, T.; Batra, I.; Luthra, M.; Vimal, S.; Dhiman, G.; Malik, A.; Shabaz, M. Gene expression-assisted cancer prediction techniques. *J. Healthc. Eng.* **2021**, *2021*, 4242646. [CrossRef] [PubMed]
2. Ahluwalia, P.; Kolhe, R.; Gahlay, G.K. The clinical relevance of gene expression based prognostic signatures in colorectal cancer. *Biochim. Biophys. Acta Rev. Cancer* **2021**, *1875*, 188513. [CrossRef] [PubMed]
3. Schaafsma, E.; Fugle, C.M.; Wang, X.; Cheng, C. Pan-cancer association of HLA gene expression with cancer prognosis and immunotherapy efficacy. *Br. J. Cancer* **2021**, *125*, 422–432. [CrossRef]
4. Tourang, M.; Fang, L.; Zhong, Y.; Suthar, R.C. Association between Human Endogenous Retrovirus K gene expression and breast cancer. *Cell. Mol. Biomed. Rep.* **2021**, *1*, 7–13. [CrossRef]
5. Satyananda, V.; Oshi, M.; Endo, I.; Takabe, K. High BRCA2 gene expression is associated with aggressive and highly proliferative breast cancer. *Ann. Surg. Oncol.* **2021**, *28*, 7356–7365. [CrossRef]
6. Qian, Y.; Daza, J.; Itzel, T.; Betge, J.; Zhan, T.; Marmé, F.; Teufel, A. Prognostic cancer gene expression signatures: Current status and challenges. *Cells* **2021**, *10*, 648. [CrossRef] [PubMed]
7. Munkácsy, G.; Santarpia, L.; Győrffy, B. Gene Expression Profiling in Early Breast Cancer—Patient Stratification Based on Molecular and Tumor Microenvironment Features. *Biomedicines* **2022**, *10*, 248. [CrossRef]
8. Oliveira, L.J.C.; Amorim, L.C.; Megid, T.B.C.; De Resende, C.A.A.; Mano, M.S. Gene expression signatures in early Breast Cancer: Better together with clinicopathological features. *Crit. Rev. Oncol. Hematol.* **2022**, *175*, 103708. [CrossRef]
9. Schettini, F.; Chic, N.; Brasó-Maristany, F.; Paré, L.; Pascual, T.; Conte, B.; Martínez-Sáez, O.; Adamo, B.; Vidal, M.; Barnadas, E.; et al. Clinical, pathological, and PAM50 gene expression features of HER2-low breast cancer. *NPJ Breast Cancer* **2021**, *7*, 1. [CrossRef]

10. Zhong, Y.; Chalise, P.; He, J. Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data. *Commun. Stat. Simul. Comput.* **2023**, *52*, 110–125. [CrossRef]

11. Petinrin, O.O.; Saeed, F.; Li, X.; Ghabban, F.; Wong, K.C. Reactions' descriptors selection and yield estimation using metaheuristic algorithms and voting ensemble. *Comput. Mater. Contin.* **2022**, *70*, 4745–4762.

12. Hameed, S.S.; Petinrin, O.O.; Hashi, A.O.; Saeed, F. Filter-wrapper combination and embedded feature selection for gene expression data. *Int. J. Adv. Soft Compu. Appl.* **2018**, *10*, 90–105.

13. Townes, F.W.; Hicks, S.C.; Aryee, M.J.; Irizarry, R.A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **2019**, *20*, 295. [CrossRef] [PubMed]

14. Jain, I.; Jain, V.K.; Jain, R. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Appl. Soft Comput.* **2018**, *62*, 203–215. [CrossRef]

15. Kabir, M.F.; Chen, T.; Ludwig, S.A. A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthc. Anal.* **2023**, *3*, 100125. [CrossRef]

16. Prasad, Y.; Biswas, K.; Hanmandlu, M. A recursive PSO scheme for gene selection in microarray data. *Appl. Soft Comput.* **2018**, *71*, 213–225. [CrossRef]

17. Sharbaf, F.V.; Mosafer, S.; Moattar, M.H. A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. *Genomics* **2016**, *107*, 231–238. [CrossRef]

18. Alhenawi, E.A.; Al-Sayyed, R.; Hudaib, A.; Mirjalili, S. Improved intelligent water drop-based hybrid feature selection method for microarray data processing. *Comput. Biol. Chem.* **2023**, *103*, 107809. [CrossRef]

19. Keshta, I.; Deshpande, P.S.; Shabaz, M.; Soni, M.; Bhadla, M.K.; Muhammed, Y. Multi-stage biomedical feature selection extraction algorithm for cancer detection. *SN Appl. Sci.* **2023**, *5*, 131. [CrossRef]

20. Sayed, S.; Nassef, M.; Badr, A.; Farag, I. A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. *Expert Syst. Appl.* **2019**, *121*, 233–243. [CrossRef]

21. Li, X.; Wang, H. On Mean-Optimal Robust Linear Discriminant Analysis. In Proceedings of the 2022 IEEE International Conference on Data Mining (ICDM), Orlando, FL, USA, 30 November–3 December 2022; pp. 1047–1052.

22. Li, X.; Wang, H. Adaptive Principal Component Analysis. In Proceedings of the 2022 SIAM International Conference on Data Mining (SDM), Alexandria, VA, USA, 28–30 April 2022; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2022; pp. 486–494.

23. Jiang, J.; Xu, J.; Liu, Y.; Song, B.; Guo, X.; Zeng, X.; Zou, Q. Dimensionality reduction and visualization of single-cell RNA-seq data with an improved deep variational autoencoder. *Briefings Bioinform.* **2023**, *24*, bbad152. [CrossRef] [PubMed]

24. Hameed, S.S.; Muhammad, F.F.; Hassan, R.; Saeed, F. Gene Selection and Classification in Microarray Datasets using a Hybrid Approach of PCC-BPSO/GA with Multi Classifiers. *J. Comput. Sci.* **2018**, *14*, 868–880. [CrossRef]

25. Dettling, M.; Bühlmann, P. Supervised clustering of genes. *Genome Biol.* **2002**, *3*, research0069.1. [CrossRef]

26. Alon, U.; Barkai, N.; Notterman, D.A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 6745–6750. [CrossRef] [PubMed]

27. Zhu, Z.; Ong, Y.S.; Dash, M. Markov Blanket-Embedded Genetic Algorithm for Gene Selection. *Pattern Recognit.* **2007**, *49*, 3236–3248. [CrossRef]

28. Microarray Datasets. Available online: https://csse.szu.edu.cn/staff/zhuzx/Datasets.html (accessed on 8 June 2023).

29. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **1999**, *286*, 531–537. [CrossRef] [PubMed]

30. Dudoit, S.; Fridlyand, J.; Speed, T.P. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **2002**, *97*, 77–87. [CrossRef]

31. Díaz-Uriarte, R.; De Andres, S.A. Gene selection and classification of microarray data using random forest. *BMC Bioinform.* **2006**, *7*, 3. [CrossRef]

32. Nguyen, H.M.; Cooper, E.W.; Kamei, K. Borderline over-sampling for imbalanced data classification. In Proceedings of the Fifth International Workshop on Computational Intelligence & Applications, IEEE SMC Hiroshima Chapter, Hiroshima, Japan, 10–12 November 2009.