Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2025;17(2):81-91

ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA DOI: 10.5336/biostatic.2025-111609

Comparison of the Effect of Dimension Reduction Methods on Classification Performance in Gene Expression Data Sets: Observational Study

Gen İfadesi Veri Setlerinde Boyut Azaltma Yöntemlerinin Sınıflandırma Performansına Etkisinin Karşılaştırılması: Gözlemsel Çalışma

Fatma Hilal YAĞIN[a], Harika Gözde GÖZÜKARA BAĞ[a]

[a]İnönü University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye

This study was presented as a summary orally in XXII. National and V. International Biostatistics Congress in October 28-30, 2021, Online.
This study was prepared based on the findings of Fatma Hilal Yagin's thesis study titled "Comparison of the effect of dimension reduction methods on classification performance in gene expression data sets" (Malatya: Inonu University; 2020).

ABSTRACT Objective: The aim of this study is to compare the effects of dimensionality reduction methods [least absolute shrinkage and selection operator (LASSO), principal component analysis (PCA), and independent component analysis (ICA)] on various support vector machine (SVM) classification methods in high-dimensional acute myeloid leukemia (AML) gene expression data. **Material and Methods:** In this study, gene expression omnibus database was used to analyze gene expression profiles in AML patients. Data included expression levels for 64 individuals and 22,283 genes. SVM with different kernel functions were used in dimensionality reduction analyses LASSO, PCA, and ICA classification analyses. 10-fold cross-validation with 10 iterations and random search were used for resampling and hyperparameter optimization. The performance of the model was evaluated using the average accuracy, sensitivity, specificity, precision, and F criterion of 500 iterations. **Results:** AML data were filtered to reveal 6,201 genes. After PCA/ICA, 10 components were extracted, and 21 genes were selected as biomarkers for AML disease. While the polynomial kernel function model with PCA achieved the highest accuracy, the SVM models with polynomial kernel function showed the best performance for all analyses. The models were then selected for their potential biomarkers. **Conclusion:** In order to build classification models using gene expression data, high dimensionality should be eliminated using dimensionality reduction methods. This reduces the analysis time and improves the prediction performance. In the AML gene expression dataset, SVM models with polynomial kernel function give better results than linear and radial basis models.

Keywords: Dimension reduction; gene expression; feature extraction; feature selection; biomarker discovery

ÖZET Amaç: Bu çalışmanın amacı, yüksek boyutlu akut miyeloid lösemi (AML) hastalığı gen ifadesi verilerinde boyut azaltma yöntemlerinin [en az mutlak küçülme ve seçim operatörü (least absolute shrinkage and selection operator "LASSO"), temel bileşen analizi (principal component analysis "PCA") ve bağımsız bileşen analizi (independent component analysis "ICA")], çeşitli destek vektör makinesi [support vector machine (SVM)] sınıflandırma yöntemleri üzerindeki etkilerini karşılaştırmaktır. **Gereç ve Yöntemler:** Bu çalışmada, AML hastalarında gen ekspresyon profillerini analiz etmek için gen ekspresyon omnibus veri tabanı kullanılmıştır. Veriler, 64 kişi ve 22.283 gen için ifade düzeylerini içermektedir. Boyut azaltma analizleri LASSO, PCA ve ICA sınıflandırma analizlerinde, farklı çekirdek fonksiyonlardaki SVM kullanıldı. Yeniden örnekleme için 10 tekrarlı 10 kat çapraz doğrulama ve hiperparametre optimizasyonu için rastgele arama kullanılmıştır. Modelin performansı, 500 tekrarlı örneğin ortalama doğruluk, duyarlılık, seçicilik, kesinlik ve F kriteri kullanılarak değerlendirilmiştir. **Bulgular:** AML verileri filtrelenerek 6.201 gen ortaya çıkarılmıştır. PCA/ICA sonrasında 10 bileşen çıkarılmış ve 21 gen, AML hastalığı için biyobelirteç olarak seçilmiştir. PCA ile polinom çekirdek fonksiyonu modeli en yüksek doğruluk elde ederken, polinom çekirdek fonksiyonlu SVM modelleri tüm analizler için en iyi performansı göstermiştir. Modeller, daha sonra potansiyel biyobelirteçleri için seçilmiştir. **Sonuç:** Gen ifadesi verilerini kullanarak sınıflandırma modelleri oluşturmak için boyut azaltma yöntemleri kullanılarak yüksek boyutluluk ortadan kaldırılmalıdır. Bu durum, analiz süresini kısaltır ve tahmin performansını artırır. AML gen ifadesi veri setinde polinomial çekirdek fonksiyonuna sahip SVM modelleri, doğrusal ve radyal tabanlı modellerden daha iyi sonuçlar vermektedir.

Anahtar kelimeler: Boyut indirgeme; gen ifadesi; özellik çıkarımı; özellik seçimi; biyobelirteç keşfi

TO CITE THIS ARTICLE:
Yağın FH, Gözükara Bağ HG. Comparison of the effect of dimension reduction methods on classification performance in gene expression data sets: Observational study. Turkiye Klinikleri J Biostat. 2025;17(2):81-91.

Correspondence: Fatma Hilal YAĞIN
İnönü University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye
E-mail: hilal.yagin@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 28 Apr 2025 ***Received in revised form:*** 16 Aug 2025 ***Accepted:*** 18 Aug 2025 ***Available online:*** 08 Sep 2025

Today, with the development of technology, the increasing data size gives rise to multidisciplinary studies and fields. One of these fields, bioinformatics, is becoming indispensable in biomedical research. There are 2 main reasons why researchers from different disciplines who work with smaller data sets turn to research fields such as bioinformatics, where much higher-dimensional data sets are available. The first of these is the Human Genome Project, which is considered a natural result of the history of genetic research, and the second is microarray technology, which allows rapid and economical measurement of gene expression levels for thousands of genes at the same time.[1,2]

With the Human Genome Project, an international research program that comprehensively examines the structure, organization and function of human genes, results were completed for 90% of the three billion base pairs of the entire genome in February 2001 and 100% in April 2003.[2,3] Research conducted through this project has led to a large and growing library of organisms containing high-dimensional data sets of sequenced genomes.[1]

Microarray technology has made it possible to analyze thousands of gene expression profiles related to many diseases, especially cancer, simultaneously. Data mining methods are gaining importance in cancer research using this technology. Classification of gene expression profiles with the help of microarray data sets is becoming a common study in biomedical research. Success in diagnosis and treatment can be achieved directly by determining "biomarker" genes that may be effective for diseases. Thanks to the determined biomarkers, personalized preventive treatments can be applied before diseases progress. For this reason, a significant part of cancer research consists of cancer classification, discovery of cancer subclasses, and selection of the most important genes that can be biomarkers for the type of cancer of interest.[4]

Gene expression datasets obtained with microarray technology generally contain a large number of gene information belonging to a small number of patients. These datasets, which can be defined as high-dimensional for data mining, reduce model performance during the modeling phase. For this purpose, dimensionality reduction analyses should be performed with feature selection and/or feature extraction methods before performing classification analyses on gene expression datasets. After selecting genes/components with distinguishing features for the disease, classification models can be created to both shorten the analysis time and increase the performance of the created models.

This study aims to investigate the effects of feature extraction and feature selection methods on the performance of the classification model for dimensionality reduction in the acute myeloid leukemia (AML) disease gene expression dataset. For this purpose, after applying the necessary preprocessing steps to the dataset; principal component analysis (PCA) and independent component analysis (ICA) from feature extraction methods, and least absolute shrinkage and selection operator (LASSO) method from feature selection methods were applied to the dataset separately. Then, classification models were established on the reduced data sets with the support vector machine (SVM) algorithm with linear, polynomial and radial based kernel functions.

MATERIAL AND METHODS

DATASET

In this study, the AML dataset uploaded to the gene expression omnibus (GEO) data repository with the code GDS3057 was used. Since the current study uses gene expression data shared as open access, ethics committee approval is not required. AML is one of the most common and fatal forms of hematopoietic malignancies, and conducted a study on the role of some genes over expressed in AML based on the hypothesis that previously unnoticed expression changes that occur only in AML blasts can be identified by microarray. To test this hypothesis, gene expression profiles were compared between normal hematopoietic cells from 38 healthy donors and leukemic blasts from 26 AML patients. The AML dataset contains expression levels of 22,283 genes from 64 individuals.[5]

METHODS

All analyses in this study were performed in R 3.6.3 program. R program can be downloaded free of charge from http://cran.rproject.org/. The reasons for preferring R program are; it has a structure that provides fast results in high-dimensional data sets such as gene expression data, and it is user-friendly. In the study, normalization function in affyPLM package was used for normalization operations, PCA function in mixOmics package for PCA, fastICA function in fastICA package for ICA, and glmnet function in glmnet package for LASSO feature selection.[6-9] PCA, ICA, LASSO feature selection, and feature extraction methods were applied to reduce the size of AML gene expression data set. Normalization process was performed before PCA method was applied to AML data set used in the study. Before feature selection and feature extraction methods were applied, nsFilter function in geneFilter package was used for gene filtering process in R program. In this function, the features that show little change between samples or constantly low signal as a result of filtering are filtered out.[10] In the modeling phase, iterative 10-fold cross-validation method was used as the resampling method. Different kernel function SVMs in the caret (Classification and Regression Training) package were applied to the reduced data sets.[11] Random search method was used for hyperparameter optimization. Correct classification rates of the created classification models were given to evaluate the model performance. In addition to the correct classification rate, sensitivity, which expresses the rate of positive test results among those who are actually sick, specificity, which expresses the rate of negative test results among those who are actually healthy, and precision, which expresses the ratio of the number of true positive samples predicted as class 1 to the total number of samples predicted as class 1, were calculated as performance measures. However, precision and sensitivity measures alone are not sufficient to draw a meaningful comparison result. For this reason, it is more correct to evaluate both measures together. In addition to the correct classification rate, sensitivity, specificity, and precision criteria, the F criterion, which is calculated as the harmonic mean of the sensitivity and precision criteria, is also given. In addition to all these criteria, the analysis times are also given in seconds in order to see the effects of dimensionality reduction analyses on the modeling time.

LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR FEATURE SELECTION

LASSO method was first used by Robert Tibshirani in 1996. The 2 main tasks of the method are regularization and feature selection. LASSO method imposes a constraint on the sum of absolute values of model parameters, the sum must be less than a fixed value (upper bound). To do this, the method applies a narrowing (regularization) process in which it penalizes the coefficients of regression features that reduce some of them to zero. During feature selection, variables that still have a non-zero coefficient after the narrowing process are selected into the model. The aim of this process is to minimize the estimation error. In practice, the parameter λ, which controls the strength of the penalty, is of great importance. When λ is large enough, the coefficients are forced to be exactly equal to zero, thus reducing dimensionality. The larger the parameter λ is, the more coefficients are reduced to zero. There are many advantages of using the LASSO method, first of all, it can provide very good estimation accuracy, because the reduction and removal of coefficients can reduce the variance without a significant increase in the bias. It is especially useful when there are few observations and many features in the data set. Also, LASSO helps to increase the interpretability of the model by eliminating irrelevant features that are not correlated with the response feature (factor), in this way the problem of overfitting can be eliminated.[12] It is considered a constrained optimisation issue that intend to make the absolute weights being less than a constant *t (ref)*, as what is expressed in the formula. The formula has y as the dependent feature, *x* as the independent feature, α as the regression model constant, and β as the coefficient of *P* independent features. The *t* on the right-hand side represents the penalty coefficient, when the $t<t_0$, some coefficients in the regression model will become 0 and be eliminated, thus achieving the effect of feature screening.

$$(\hat{\alpha},\hat{\beta})= \arg\ \min \{\Sigma_{i=1}^{N}[y_i - \alpha - \Sigma_{j=1}^{\rho}\beta_i x_{ij}]^2\},\ \ S.T.\Sigma_{i=1}^{\rho}|\beta_i| \leq t \quad (1)$$

PRINCIPAL COMPONENT ANALYSIS

PCA performs transformations on a dataset (with dimensions D) to create a new coordinate system (with a new set of dimensions, D'). The variance in the data is such that D′1 varies more than D′2, while D′2 varies more than D′3. In simpler terms, the most important information about the data is fit into one dimension, then the same is done for a second dimension, and so on, until the same number of created dimensions are found in the data as were originally present. This means that the first dimension can describe the original structure of the data more than the second. As a dimensionality reduction method, in PCA, less important dimensions can be omitted to describe the data. For example, if the user specifies that 95% of the variance should be preserved, D′ will contain the smallest set of dimensions that preserves at least 95% of the original variance. When focusing specifically on how the theoretical basis of an algorithm affects performance, one should consider the metric by which the dimensions are determined (i.e. variance). When data are continuous, multivariate normal, and independent, variance will be a useful metric. In addition, data sets that do not meet these assumptions will give poor results when analyzed with PCA. Therefore, it is essential to check whether the assumptions are met before applying PCA. In addition, PCA is a method that finds linear components in the data. However, nonlinear components can also be present in the data. Running PCA on such data may yield worse results than algorithms that do not make such an assumption.[13]

INDEPENDENT COMPONENT ANALYSIS

The aim of ICA is to obtain a linear transformation of non-normally distributed data in a way that will make them statistically independent or as independent as possible. ICA was first put forward by Hérault et al., in a study on the development of a simplified model of the movement in muscle contraction.[14] The name independent components was first mentioned in an article written by Comon.[15] ICA has a wide range of applications in different disciplines such as genetics, image processing, brain tomography, communication, finance, seismology, etc. In ICA, it is assumed that multivariate data consists of a linear combination of a number of independent components (factors). The number of components is generally taken to be equal to the number of variables. If we represent the data set consisting of p variables, each sampled at n points, with the Z matrix, the Z matrix in the ICA model is

$$Z = AY$$

is expressed by matrix multiplication. In this equation, A: represents the mixture matrix, Y: represents the source matrix containing the independent components. Here, both the mixture matrix and the source matrix are unknown. Under ICA, both matrices are estimated using only the original data matrix. First, the mixture matrix is estimated. Then, the matrix containing the independent components is obtained by multiplying the inverse of A with the Z data matrix. In order for the ICA model to be defined, the independent components must not be normally distributed. In addition, it is assumed that the number of mixtures is equal to the number of independent components. This assumption facilitates the operations required for separation. ICA is similar to multivariate methods such as PCA, factor analysis (FA), and minimum/maximum autocorrelation factors (MOF) analysis. Multivariate data sets are expressed as a linear combination of factors in all of these methods. While uncorrelated and normally distributed factors are obtained with PCA, FA, and MOF methods, when the data set is analyzed with ICA, independent and also non-normally distributed factors are obtained.[16,17]

SUPPORT VECTOR MACHINE MODELS

SVM is one of the classification methods that has been frequently used in recent years and gives quite good results. This method is used in different areas such as diagnosis of diseases, different engineering applications, etc.[18] The SVM method proposed by Cortes and Vapnik uses the principle of minimizing

structural risk. In this technique, a plane is examined in which the distances of 2 classes/groups to the closest observations are maximized.[19] Data that are not classified with linear methods are mapped to a larger dimensional space by the SVM method using a nonlinear function. A second-order optimization equation can be used in training SVM.[20] At the same time, in machine learning methods, SVMs (also called support vector networks) are models supervised by learning algorithms used for classification and regression analysis. Given a set of training examples, each of which is marked as belonging to one or the other of 2 categories, an SVM training algorithm creates a model that assigns new examples to 1 or the other of the examined categories using a binary classifier. In an SVM model, examples are represented as points in space, and examples of separate categories are divided by a gap as open as possible. New examples are then assigned to the same space, and a category is predicted based on which side of the gap the new examples belong to. In addition to linear classification, SVMs can also successfully handle nonlinear classification problems, implicitly mapping their inputs to high-dimensional feature spaces. When the possible groups to which data can be assigned are not known, supervised learning is not possible. In this case, an unsupervised learning approach is required that performs the natural clustering of data into groups and then assigns new data to these groups.[21] The soft-margin SVM solves the following optimization problem:

$$\min_{w,b,\xi} \; \tfrac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi i$$

subject to:

$$yi(w^T \emptyset(xi) + b) \geq 1 - \xi i, \quad \xi i \geq 0$$

In this study, various SVM models were constructed with different kernel functions, as detailed in, and applied to the dimensionally reduced datasets using PCA, ICA, and LASSO methods (Table 1).

HYPERPARAMETER OPTIMIZATION

The random search method, which is one of the frequently used methods in the literature for hyperparameter optimization, was used in the study. The method was first proposed by Bergstra and Bengio in an article published in 2012. In the random search method, hyperparameter ranges are determined using preliminary information about the problem. Then, instead of trying each of the values in this range, hyperparameter groups are created by selecting random values. The model is trained with different random parameter groups until the parameters that provide the best performance are found.[22]

TABLE 1: Kernel types and functions

Kernel types	Function
Linear	$k(x,y)= x^T y + c$
Polynomial	$k(x,y)=(ax^T y + c)^d$
Radial basis	$k(x,y)= \exp(-\gamma \|x-y\|^2$

RESULT

After filtering the AML dataset containing 22,283 gene expression data with Nsfilter, 6,201 genes remained in the dataset. After applying PCA, ICA, LASSO methods to the filtered dataset depending on the purpose of the study, linear, polynomial and radial based kernel function SVM methods were applied. The 10 principal components created with PCA explained 89% of the total variance in the dataset. The results for the training dataset for each method used are given (Table 2). The correct classification rates of the linear, polynomial and radial based SVM models created with 21 genes selected after LASSO feature selection from the models established for the training data set were found to be 1 for the training set. The accuracy rates of the polynomial and radial based SVM models created with 10 selected components after ICA were also found to be 1 for the training set.

TABLE 2: Results of the models for the training dataset

Dimensionality reduction method	Optimization parameters	Parameter values	Accuracy (%)
Linear	C	4	98.52
Linear/PCA	C	323.29	99.69
Linear/ICA	C	0.66	99.83
Linear/LASSO	C	0.052	1
Polynomial	C	188.94	99.03
	Degree	1	
	Scale	0.0001	
Polynomial/PCA	C	1.804	99.85
	Degree	2	
	Scale	0.0048	
Polynomial/ICA	C	0.38	1
	Degree	3	
	Scale	1.89	
Polynomial/LASSO	C	0.067	1
	Degree	1	
	Scale	0.067	
Radial based	C	2.01	99.13
	sigma	7.85	
Radial based/PCA	C	0.20	99.23
	sigma	0.022	
Radial based/ICA	C	0.72	1
	sigma	0.02	
Radial based/LASSO	C	0.03	1
	sigma	0.031	

PCA: Principal component analysis; ICA: Independent component analysis; LASSO: Least absolute shrinkage and selection operator

Considering the accuracy value, the highest values in classifying the AML gene expression data set are in the models established with Polynomial SVM. The accuracy rates were found as; PCA+polynomial (95.64%), ICA+polynomial (95.41%), LASSO+polynomial (95.38%), respectively (Table 3).

TABLE 3: Results of the models for the test data set

Feature extraction/ selection method	SVM kernel function	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-score (%)	Analysis time (sec.)
Original data	Linear	94.10	93.22	95.25	94.79	93.31	2,155.8
	Polynomial	95.05	94.24	96.04	95.08	94.08	2,174.77
	Radial based	95.01	94.35	95.56	94.58	93.92	2,252.36
PCA	Linear	95.20	93.72	96.22	95.28	93.97	35.78
	Polynomial	95.64	94.53	96.46	95.48	94.99	36.33
	Radial based	94.94	93.54	95.90	94.91	93.65	36.07
ICA	Linear	95.15	93.83	96.05	95.09	93.91	34.24
	Polynomial	95.41	94.02	96.03	95.1	94.05	38.97
	Radial based	95.07	93.84	95.91	94.94	93.86	35.74
LASSO	Linear	94.98	93.71	95.85	94.90	93.75	32.67
	Polynomial	95.38	93.36	96.27	95.39	93.84	33.54
	Radial based	95.23	94.26	95.89	94.90	94.05	35.53

SVM: Support vector machine; PCA: Principal component analysis; ICA: Independent component analysis; LASSO: Least absolute shrinkage and selection operator

When examined, the gene that contributes the most to the model out of 21 genes obtained with the LASSO feature selection method is the gene with probe number 219624_at (Figure 1).

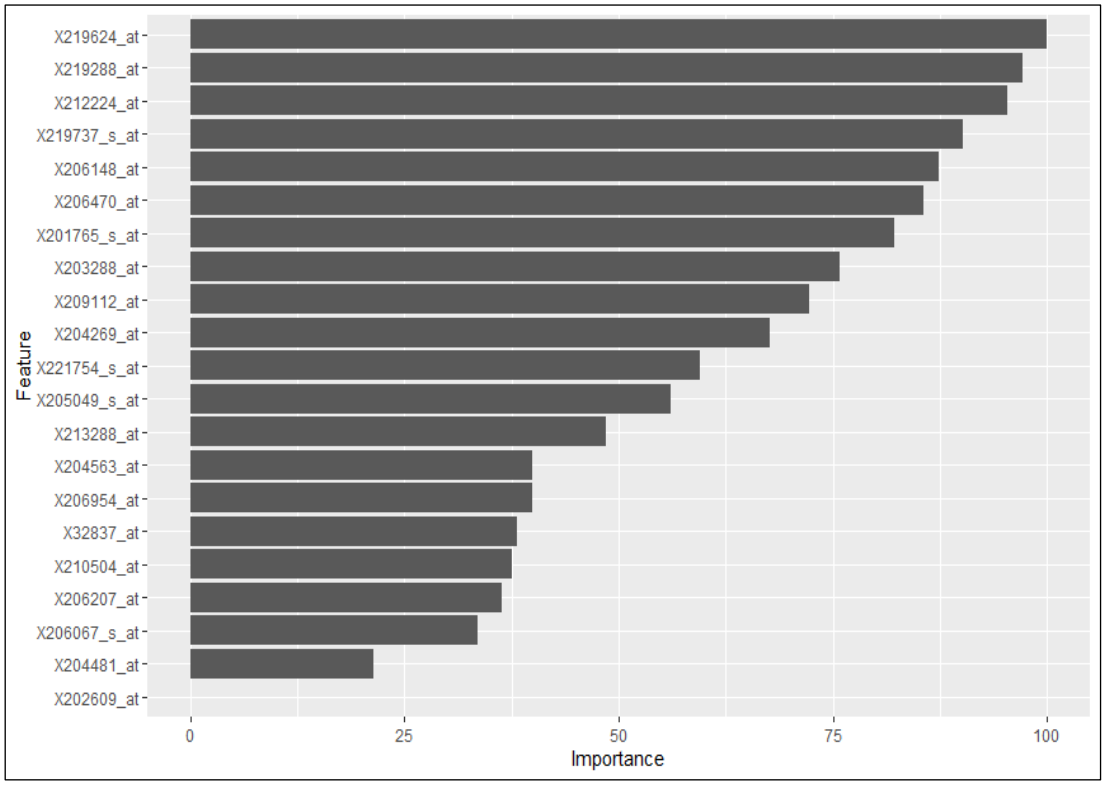

FIGURE 1: Importance order of genes added to the model after LASSO
LASSO: Least absolute shrinkage and selection operator

Information about some of the 21 genes selected with the LASSO feature selection method was obtained from the BioGPS database and is given (Table 4). When Table 4 is examined, the genes selected with the LASSO method are effective genes in the diagnosis and treatment of AML disease and are genes that can be biomarkers in AML disease.

TABLE 4: Descriptions for some of the genes selected by LASSO feature selection

Probe set name	Descriptive	Gene symbol	Function
219624_at	BAG cochaperone 4	*BAG4*	The protein encoded by this gene is a member of the *BAG* gene family of related proteins and is an anti-apoptotic protein (anti-apoptotic protein: cell proliferating protein). The *BAG* gene family is expressed at low levels in normal blood cells but is highly expressed in both primary leukemia and established cell lines of leukemia patients. *BAG* gene expression levels are higher in drug-resistant patients compared to chemotherapy-sensitive patients.[23]
219288_at	chromosome 3 open reading frame 14	*C3orf14*	The *C3orf14* gene, found in the study conducted for the discovery and validation of expression data of the Washington University Acute Myeloid Leukemia Program Genomics, was found to be a gene that could be a biomarker for AML disease.[24]
212224_at	aldehyde dehydrogenase 1 family member A1	*ALDH1A1*	Leukemic cells with toxic *ALDH1A1* substrates may be a novel targeted therapeutic strategy for AMLs.[25]
219737_s_at	protocadherin 9	*PCDH9*	*PCDH9* upregulation is a poor prognostic factor in ALL. That is, it is highly expressed in leukemia.[26]
206148_at	interleukin 3 receptor subunit alpha	*IL3RA*	The *IL3RA* gene shows increased expression in AML samples compared to normals.[5]
206470_at	plexin C1	*PLXNC1*	*PLXNC1* gene expression is decreased in AML.[5]
201765_s_at	hexosaminidase subunit alpha	*HEXA*	There is information in the literature that Tay-Sachs disease is caused by mutation in the alpha subunit of hexosaminidase.[27]
209112_at	cyclin dependent kinase inhibitor 1B	*CDKN1B*	*CDKN1B* is a potential candidate gene for prognosis in AML.[28]
204269_at	Pim-2 proto-oncogene, serine/threonine kinase	*PIM2*	It may be a treatment target for AML patients.[29]
206207_at	Charcot-Leyden crystal galectin	*CLC*	May be associated with myeloid leukemias.[30]

AML: Acute myeloid leukemia; ALL: Acute lymphoblastic leukemia

DISCUSSION

The dimensions of gene expression data sets obtained from microarray experiments are quite high. Due to their high dimensionality, modeling analyses with gene expression data sets take a long time and therefore these data sets can lead to computational inefficiency in analyses. The high dimensionality problem can also cause a decrease in model performance. In addition, a large number of genes in gene expression data sets can cause a classification algorithm that overfit to the training examples and generalize new examples poorly. In order to eliminate these problems, the results of two different forms of dimensionality reduction; PCA, ICA, LASSO methods from feature selection and feature extraction methods were comparatively examined in this study. According to the results, models created with the raw data set without applying dimensionality reduction methods caused computational inefficiency for the AML gene expression data set. When the analysis times were examined, the model established with PCA was completed in the shortest time (35.78 sec.), and the model established with radial-based SVM without any processing to the data set was completed in the

longest time (2,252.36 sec.). In addition, the classification models (linear, polynomial and radial based SVM) established without dimensionality reduction for both training and test data sets showed lower performance compared to the models established after dimensionality reduction analyses. In other words, feature selection and feature extraction methods applied before classifying the AML gene expression data set increased the classification performance. When the PCA and ICA feature extraction methods are compared, PCA performed better in the models established with linear SVM and polynomial kernel function SVM for the test data set, and ICA performed better in the SVM model established with radial basis function. When the models established as a result of LASSO feature selection and PCA/ICA feature extraction methods are examined; the model established with radial based SVM after LASSO feature selection method showed better performance than the models established with radial based SVM after PCA/ICA feature extraction methods. The best SVM models established with all dimensionality reduced data sets as a result of PCA, ICA and LASSO methods are the models established with polynomial kernel function. In addition, 21 biomarker genes most related to AML disease were selected as a result of LASSO feature selection. Probe numbers of these selected genes were examined from BioGPS database and literature. There is information in the literature that these selected genes can be used as biomarker genes in diagnosis and treatment of AML disease. These genes selected with LASSO method can be used for networks that will help to detect other genes that cause AML disease. *PIM2* and *CDKN1B* genes, which are among the genes selected with LASSO feature selection method, are genes that can be treatment targets for AML patients. Drugs that will be developed targeting this gene may be effective in the treatment of AML disease.

When compared to models that were processed with dimensionality reduction, models that were trained on unprocessed data had lower accuracy and required significantly longer computation time - as was displayed by a 2,252.36 second run of radial-based SVM vs only 35.78 seconds for PCA with linear SVM. The findings are in line with new research that highlights the important role of dimensionality reduction in biological data analysis, especially on high dimensions. Just an example is the application of machine learning in genomic data analysis, which has greatly enhanced precision medicine by enhancing diagnostic precision, and making clinical decisions more informed. Remarkably, deep learning models have routinely produced better accuracy and F1 scores than conventional approaches and as such they can potentially add value in precision medicine.[31] Despite these strengths, our study has limitations. The sample size (n=64)--though typical for genomic studies--restricts broad generalizability. While rigorous resampling (10-fold cross validation with 500 iterations) minimized overfitting risks, validation in larger, independent cohorts (e.g., TCGA-AML) is essential. Furthermore, external validation of the 21 LASSO-selected biomarkers is underway to confirm their clinical utility across diverse populations.

Both PCA and ICA are approaches for deriving new components from the data whereby they are different from the role played by LASSO in selecting particular, relevant genes for AML classification. In linear and polynomial kernel SVM models, PCA outperformed ICA, with much weight attributed to effective variance extraction in linearly organized data in a comprehensive review - research publication.[32] On the contrary, the ICA showed a modest advantage with radial basis function SVM, which suggests the capability of uncovering statistically independent features might be more in line with the non-linear assumptions of such models. Selecting 21 genes, LASSO was effective at dimensionality reduction as well as enabling the identification of biomarkers necessary for precision medicine. An examination of AML's genetic profile affirms the clinical desirability of *PIM2*, *CDKN1B* and other LASSO chosen genes as potential targets for therapy, highlighting their importance in biology.[33] The superior performance of polynomial kernel SVM across all dimensionality-reduced datasets suggests that this kernel effectively captures complex, non-linear relationships in gene expression profiles. This observation is consistent with the study, which demonstrated that kernel choice significantly impacts classification outcomes in high-dimensional settings.[34] The computational efficiency gains-evident in reduced analysis times-further underscore the practical utility of these methods in clinical bioinformatics, where rapid processing is essential for real-time diagnostics. While PCA/ICA/LASSO provided interpretable feature reduction, emerging deep learning methods (e.g., varia-

tional autoencoders) may better capture nonlinear structures in gene expression data. Future work will compare these approaches to optimize the trade-off between computational efficiency and biological insight.

CONCLUSION

For this study, AML gene expression dataset was analyzed utilizing three dimensional reduction methods (LASSO, PCA, ICA) with SVM classifiers. Notable findings reveal the serious improvement of computational speed and accuracy of models, with the data dimensionality reduced significantly. The polynomial kernel SVM outperformed others over and above PCA with its score of 95.6% accuracy. 21 genes were identified by LASSO, including the known AML biomarkers *PIM2* and *CDKNILB*, as well as *HEXA*, a gene previously connected to Tay-Sachs disease. Such results underline the necessity of preprocessing high-dimensional genomic data while suggesting *HEXA* as a new candidate for future AML research. Additional research should confirm these results on different datasets and discuss the possible involvement of *HEXA* in leukemia pathogenesis. Dimensionality reduction (LASSO/PCA/ICA) significantly improved AML classification efficiency and accuracy using SVM, with polynomial kernels achieving optimal performance (95.64% accuracy post-PCA). We identified 21 candidate biomarkers via LASSO, including established targets (*PIM2*, *CDKN1B*) and novel candidates (*HEXA*).

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

All authors contributed equally while this study preparing.

REFERENCES

1. Dziuda DM. Data Mining for Genomics And Proteomics: Analysis Of Gene And Protein Expression Data. Vol. 1. 1st ed. New Jersey: John Wiley & Sons; 2010. [Crossref] [PubMed]
2. Apitz JC. A statistical method for selection, classification, and network construction in genetic systems [Master thesis]. USA, CA: California State University; 2016. [Link]
3. Başaran E, Aras S, Cansaran-Duman D. General outlook and applications of genomics, proteomics and metabolomics. Turk Hij Den Biyol Derg. 2010;67(2):85-96. [Link]
4. Coşkun E, Karaağaoğlu E. Veri madenciliği yöntemleri ile mikrodizilim gen ifade analizi. Hacettepe Tıp Dergisi. 2011;42:180-9. [Link]
5. Stirewalt DL, Meshinchi S, Kopecky KJ, Fan W, Pogosova-Agadjanyan EL, Engel JH, et al. Identification of genes with abnormal expression changes in acute myeloid leukemia. Genes Chromosomes Cancer. 2008;47(1):8-20. [PubMed]
6. Bolstad B, Bolstad MB. affyPLM: Model based quality assessment of Affymetrix GeneChip data [Internet]. Bioconductor; 2013. Available from: [Link]
7. Lê Cao KA, Rohart F, Gonzalez I, Singh A. mixOmics: an R package for 'omics feature selection and multiple data integration [Internet]. Bioconductor; 2017 [Link]
8. Marchini JL, Heaton C, Ripley BD, Ripley MB. fastICA: FastICA algorithms to perform ICA and projection pursuit [Internet]. R package version 1.1-9; 2007 [cited 2025 Aug 29]. Available from: [Link]
9. Friedman J, Hastie T, Tibshirani R, Narasimhan B, Tay K, Simon N, et al. glmnet: Lasso and elastic-net regularized generalized linear models. 2009. [Link]
10. Gentleman R, Carey VJ, Huber W, Hahne F. Genefilter: methods for filtering genes from microarray experiments. 2011. [Link]
11. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, et al. The caret package. Vienna, Austria. 2012. Available from: [Link]
12. Fonti V, Belitser E. Feature selection using LASSO. 2017:1-25. [Link]
13. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and Intelligent Laboratory Systems. 1987;2(1-3):37-52. [Link]
14. Hérault J, Jutten C, Ans B. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. Proc. 10th Colloque GRETSI sur le traitement du signal et des images; 1985. p. 1017-22.Available from: [Link]

15. Comon P. Independent component analysis, a new concept? Signal Process. 1994;36(3):287-314. https:// [Crossref]
16. Chao S, Lihui C. Feature dimension reduction for microarray data analysis using locally linear embedding. In: Wong L, Chen PY, editors. Proceedings of the 3rd Asia-Pacific Bioinformatics Conference; 2005 Jan 17-21; Singapore. Singapore: Institute for Infocomm Research/World Scientific; 2005. p. 211-17. [Crossref]
17. Ehler M, Rajapakse VN, Zeeberg BR, Brooks BP, Brown J, Czaja W, et al. Nonlinear gene cluster analysis with labeling for microarray gene expression data in organ development. BMC Proc. 2011;5(Suppl 2):S3. from: https:// [Crossref]
18. Amsterdam EA, Wenger NK, Brindis RG, Casey DE Jr, Ganiats TG, Holmes DR Jr, et al. 2014 AHA/ACC Guideline for the Management of Patients with Non-ST-Elevation Acute Coronary Syndromes: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol. 2014;64(24):e139-e228. Erratum in: J Am Coll Cardiol. 2014;64(24):2713-4. Dosage error in article text. [PubMed]
19. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297. from: https:// [Crossref]
20. Eskidere Ö. A comparison of feature selection methods for diagnosis of Parkinson's disease from vocal measurements. Sigma J Eng Nat Sci. 2012;30(4):402-14. [Link]
21. Ben-Hur A, Horn D, Siegelmann HT, Vapnik V. Support vector clustering. Journal of Machine Learning Research. 2001;2:125-37. [Link]
22. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. Journal of Machine Learning Research. 2012;13:281-305. [Link]
23. Dyer JO, Dutta A, Gogol M, Weake VM, Dialynas G, Wu X, et al. Myeloid leukemia factor acts in a chaperone complex to regulate transcription factor stability and gene expression. J Mol Biol. 2017;429(13):2093-107. ; [Crossref] [PubMed] [PMC]
24. Tomasson MH, Xiang Z, Walgren R, Zhao Y, Kasai Y, Miner T, et al. Somatic mutations and germline sequence variants in the expressed tyrosine kinase genes of patients with de novo acute myeloid leukemia. Blood. 2008;111(9):4797-808. ; [Crossref] [PubMed] [PMC]
25. Gasparetto M, Pei S, Minhajuddin M, Khan N, Pollyea DA, Myers JR, et al. Targeted therapy for a subset of acute myeloid leukemias that lack expression of aldehyde dehydrogenase 1A1. Haematologica. 2017;102(6):1054-65. ; [Crossref] [PubMed] [PMC]
26. Silveira VS, Scrideli CA, Moreno DA, Yunes JA, Queiroz RG, Toledo SC, et al. Gene expression pattern contributing to prognostic factors in childhood acute lymphoblastic leukemia. Leuk Lymphoma. 2013;54(2):310-4. [Crossref] [PubMed]
27. Beutler E, Kuhl W, Comings D. Hexosaminidase isozyme in type O Gm2 gangliosidosis (Sandhoff-Jatzkewitz disease). Am J Hum Genet. 1975;27(5):628-38. ; [PubMed] [PMC]
28. Haferlach C, Bacher U, Kohlmann A, Schindela S, Alpermann T, Kern W, et al. CDKN1B, encoding the cyclin-dependent kinase inhibitor 1B (p27), is located in the minimally deleted region of 12p abnormalities in myeloid malignancies and its low expression is a favorable prognostic marker in acute myeloid leukemia. Haematologica. 2011;96(6):829-36. ; [Crossref] [PubMed] [PMC]
29. Kapelko-Slowik K, Owczarek TB, Grzymajlo K, Urbaniak-Kujda D, Jazwiec B, Slowik M, et al. Elevated PIM2 gene expression is associated with poor survival of patients with acute myeloid leukemia. Leuk Lymphoma. 2016;57(9):2140-9. [PubMed]
30. Dvorak AM, Letourneau L, Weller PF, Ackerman SJ. Ultrastructural localization of Charcot-Leyden crystal protein (lysophospholipase) to intracytoplasmic crystals in tumor cells of primary solid and papillary epithelial neoplasm of the pancreas. Lab Invest. 1990;62(5):608-15. [PubMed]
31. Sasikala R, Deepthi KJ, Balakrishnan TS, Krishnan P, Ebenezar US. Machine Learning-Enhanced Analysis of Genomic Data for Precision Medicine. In: Proceedings of the 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0; 2024 Jun 5-7; Raigarh, India. p. 1-5. [Link]
32. van der Maaten L, Postma E, van den Herik J. Dimensionality reduction: a comparative review. J Mach Learn Res. 2009;10:66–71. Available from: [Link]
33. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood. 2017;129(4):424-47. ; [Crossref] [PubMed] [PMC]
34. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning. 2002;46:389-422. [Crossref]