Middle Black Sea Journal of Health Science / Mid Blac Sea J Health Sci **December 2021; 7(3):390-396**

RESEARCH ARTICLE **DOI:** 10.19127/mbsjohs.994625

The Effects of Variable Selection and Dimension Reduction Methods on the Classification Model in the Small Round Blue Cell Tumor Dataset

Fatma Hilal Yagin[1](ID), Zeynep Kucukkakcali[1](ID), Ipek Balikci Cicek[1](ID), Harika Gozde Gozukara Bag[1] (ID)

[1]Department of Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, Malatya, Turkey.

Received: 13 September 2021, Accepted: 24 November 2021, Published online: 31 December 2021

Abstract

Objective: The purpose of this study is to investigate and compare the effects of different dimension reduction methods (PCA, ICA, PCA + Forward Selection, ICA + Forward Selection) on the K-NN classifier using open access gene expression data of small round blue cell tumor types.

Methods: In this study, open access gene expression data of small round blue cell tumor types was used for investigate and compare the effects of different dimension reduction methods. In the study, PCA, ICA, PCA + Forward Selection, ICA + Forward Selection were used as different dimension reduction methods together with K-NN classification method.

Results: Accuracy values obtained from the dimension reduction model made with PCA on K-NN model; for EWS, BL, NB, and RMS type tumors with 93.51%, 91.14%, 92.31%, and 94.74% respectively. Accuracy values obtained from the dimension reduction model made with PCA + Forward Selection on K-NN model; for EWS, BL, NB, and RMS type tumors with 96.25%, 96.25%, 95.06% and 95.47%, respectively. Accuracy values obtained from the dimension reduction model made with ICA on K-NN model; for EWS, BL, NB, and RMS type tumors with 91.89%, 90.67%, 88.31% and 89.47% respectively. Accuracy values obtained from the dimension reduction model made with ICA+ Forward Selection on K-NN model; for EWS, BL, NB, and RMS type tumors with 93.51%, 91.14%, 92.31% and 94.74% respectively.

Conclusion: In this study, the model created with PCA gives higher results than the model created with ICA. In addition, according to the results of the models obtained by applying the Forward selection method on these 2 models, the forward selection method has increased the classification performance.

Key words: Dimension reduction, principal component analysis, independent component analysis, K-NN, Small round blue cell tumor.

Suggested Citation: Yagin F H, Kucukkakcali Z, Balikci Cicek I., Gozukara Bag H.G. The Effects of Variable Selection And Dimension Reduction Methods On The Classification Model In The Small Round Blue Cell Tumor Dataset Mid Blac Sea Journal of Health Sci, 2021; 7(3):390-396

Address for correspondence/reprints: **E-mail:** zeynep.tunc@inonu.edu.tr

Zeynep Kucukakcali

Telephone number: +90 (422) 341 06 60-1337

Introduction

Small round blue cell tumor was first described by Gerald and Rosai in 1989 (1). Although its histogenesis is not known exactly, it is thought to originate from the progenitor cell with multiphenotypic differentiation potential (2). Small round blue cell tumor involves the abdominal and pelvic peritoneum diffuse and is usually observed in childhood and young adulthood. Tumors detected more in men show a rather aggressive clinical course. Although it is composed of slightly differentiated round cells, small round blue cell tumor can be differentiated from other primitive round cell tumors with morphological, immunohistochemical and genetic findings (3).

Small round blue cell tumors are four different childhood tumors with similar appearances, making accurate clinical diagnosis extremely difficult and difficult to distinguish. However, accurate diagnosis is important because treatment options, response to treatment, and prognoses differ greatly depending on the diagnosis. These include Ewing's tumors (EWS), neuroblastoma (NB), non-Hodgkin lymphoma (Burkitt lymphoma, BL), and rhabdomyosarcoma (RMS) (4).

Bioinformatics is a research field that includes the use of various computer-aided methods, retrieves and stores data to solve biological problems, and interprets data with the help of statistical analysis. Within the framework of these purposes, one of the fields of study of bioinformatics is DNA microarrays. DNA microarray technology is a method used to discover the functions of gene functions. In order to obtain more efficient results when working with microarrays, it is necessary to use an effective and robust dimension reduction algorithm to reduce the increasing feature size of the data. Dimension reduction algorithms contribute to reducing computation time in high dimensional data such as microarrays, improving prediction performance obtained by machine learning methods, and facilitating the interpretation of results (5).

Principal Component Analysis (PCA), one of these methods, is a multivariate technique that analyzes a data table in which observations are defined by several interrelated dependent variables. In other words, PCA is a transformation technique that enables the size of the data set containing a large number of related variables to be processed in a smaller size by protecting the variables in the data set (6). Independent Component Analysis (ICA) is a new method for finding the linear, orthogonal (not mandatory) coordinate system in a multivariable data set. It makes the projections of the input data on the existing coordinate system independent from each other and minimizes the relationship between them. The purpose of the ICA is to make a linear transformation that reduces the relationship between resources (7).

Forward selection, one of the frequently used methods for feature selection in gene expression data sets, is an iterative method. With each iteration, the most contributing feature is added to the model until adding a new feature does not improve the performance of the model. It is a method that creates a feature subset by testing the effect of the feature with a classifier at each step (8).

The K-nearest neighbor algorithm (K-NN) is widely used in classification due to its simple and easy implementation and the powerful and useful learning process. The K-NN method is one of the supervised learning algorithms used in classification problems and calculates the proximity of the test sample to the samples in the training set according to a predetermined distance criterion. After this process, it determines the k closest samples and includes the test sample to the class which these samples belong to the most (9).

The aim of this study is to investigate and compare the effects of different dimension reduction methods (PCA, ICA, PCA + Forward Selection, ICA + Forward Selection) on the K-NN classifier using open access gene expression data of small round blue cell tumor types.

Methods

Dataset

In this study, the Small Round Blue Cell Tumor data set was examined inclinometers (10). Small Round Blue Cell Tumor data set consists of gene expressions of four different pediatric tumors. The data set was created to facilitate tumor diagnosis by using gene expressions. Accurate clinical diagnosis is difficult due to the similar appearance of the tumors in their histology. Due to the treatment options, it is important to make the correct diagnosis, as the responses in terms of treatment and prognosis vary depending on the common diagnosis. The data set consists of 2308 variables and 83 observations. The dependent variable classes in the data set are given in Table 1.

Table 1. Distribution of the dependent variable

Dependent Variable	Count (%)
Ewing sarkomu (EWS)	29 (34.9)
Burkitt lenfoma (BL)	11 (13.3)
Nöroblastom (NB):	18 (21.7)
Rabdomiyosarkom (RMS)	25 (30.1)

In the study, firstly, the K-NN classifier was applied to the data set whose dimensions were reduced separately by PCA and ICA. A 10-fold cross validation method was used in the training and testing phase. The purpose of this process is to obtain more reliable results from the created model. Then, in order to compare the results; After the forward selection method, one of the feature selection methods, was applied to the data set whose size was reduced by PCA and ICA, the models was created with the K-NN classifier. The classification performance of the models was evaluated using Accuracy, Precision, Sensitivity, Specificity, F1-score, Matthews's correlation coefficient (MCC), and G-mean criteria.

PCA and ICA

Dimension reduction is used in a variety of computer science fields, including computer vision, pattern recognition, and machine learning. The advantages of dimension reduction are as follows: first, it typically enables the whole method to be implemented in a more computationally efficient manner. Second, it typically results in an improvement in the method's accuracy or right amount (11).

In this paper, we propose to exploit principal components analysis (PCA) and independent component analysis (ICA) for dimension reduction. PCA is commonly used in image processing, pattern recognition, data compression, data mining, machine learning, and computer vision, among other fields (12). In data mining, principal component analysis (PCA) is commonly used to investigate data structure. By maximizing the variance of the data, new orthogonal variables (latent variables or principal components) are obtained in PCA. The number of latent variables (factors) is significantly smaller than the number of original variables, allowing the data to be visualized in a low-dimensional PC space. Although PCA reduces the dimensionality of the space, it does not reduce the number of original variables because it generates new latent variables using all of the original variables (principal components). Reducing the number of variables is often beneficial for interpretation or potential inquiries (13).

Independent component analysis (ICA) is a multiple statistical method which seks to uncover disguised variables in high-dimensional data. ICA, which is a statistical computational method, is employed to find underlying hidden factors among a set of random vectors. The main aim of ICA method is to obtain the independent components (ICs), which are linearly independent or as independent as possible. In this way, ICA can be seen as a extension of Principal Components Analysis (PCA). ICA, on the other hand, is founded on statistical independence rather than unrelatedness, which is a much stronger function than unrelatedness (14).

Forward Selection

The first variable chosen for inclusion in the built model in forward selection is the one with the highest association with the dependent variable. After the variable has been chosen, it is assessed using a set of parameters. Mallows' Cp and Akaike's knowledge criterion are two of the most popular. If the first variable chosen meets the inclusion criteria, the forward selection process begins, with the statistics for variables not in the equation being used to choose the next one. When there are no more variables that meet the entry criteria, the process ends.

K-NN

The k-Nearest-Neighbors (K-NN) classification method is a non-parametric classification method that is easy to use but useful in many situations. To classify a data record t, its k closest neighbors are retrieved, and this forms a neighbourhood of t. The classification for t is typically decided by majority voting among data records in the neighborhood, with or without consideration of distance-based weighting. However, in order to use K-NN, we must select an acceptable value for k, and the classification output is highly dependent on this value. In certain ways, the K-NN approach is influenced by k. There are many methods for determining the k value, but one of the most straightforward is to run the algorithm several times with various k values and choose the one that performs best (15).

Results

Accuracy values obtained from the dimension reduction model made with PCA on K-NN model; for EWS, BL, NB, and RMS type tumors with 93.51%, 91.14%, 92.31%, and 94.74% respectively. In this model, the highest precision value was obtained in the EWS tumor type subclass, the highest specificity value in the NB tumor type subclass, and the highest sensitivity value, F1-score value, MCC value, and G-mean value were obtained from the RMS tumor type subclass (Table 2).

Table 2. PCA+ K-NN Performance metric values calculated from created models in the testing stage

Srbct tumor / Metrics	Ewing sarkomu (EWS)	Burkitt lenfoma (BL)	Nörobl astom (NB)	Rabdomiyo sarkom (RMS)
Accuracy	0.9351	0.9114	0.9231	0.9474
Precision	0.8966	0.7273	0.8889	0.8800
Sensitivity	0.9286	0.6667	0.8000	0.9565
Specificity	0.9388	0.9552	0.9655	0.9434
F1-score	0.9123	0.6957	0.8421	0.9167
Matthew's correlation coefficient (MCC)	0.8611	0.6447	0.7934	0.8799
G-mean	0.9337	0.7980	0.8789	0.9499

Accuracy values obtained from the dimension reduction model made with PCA + Forward Selection on K-NN model; for EWS, BL, NB, and RMS type tumors with 96.25%, 96.25%, 95.06% and 95.47%, respectively. In this model, the highest precision value was obtained in the EWS type tumor subclass, the highest specificity value was obtained in the NB type tumor subclass, and the highest sensitivity value, F1-score value, MCC value and G-mean value were obtained from the RMS type tumor subclass (Table 3).

Table 3. PCA+ Forward Selection+ K-NN Performance metric values calculated from created models in the testing stage

Srbct tumor / Metrics	Ewing sarkomu (EWS)	Burkitt lenfoma (BL)	Nöroblas tom (NB)	Rabdomiyo sarkom (RMS)
Accuracy	0.9625	0.9625	0.9506	0.9747
Precision	0.9310	0.8182	0.9444	0.9600
Sensitivity	0.9643	0.900	0.8500	0.9600
Specificity	0.9615	0.9714	0.9836	0.9815
F1-score	0.9474	0.8571	0.8947	0.9600
Matthews correlation coefficient (MCC)	0.9186	0.8369	0.8646	0.9415
G-mean	0.9629	0.9350	0.9144	0.9707

Accuracy values obtained from the dimension reduction model made with ICA on K-NN model; for EWS, BL, NB, and RMS type tumors with 91.89%, 90.67%, 88.31% and 89.47% respectively. In this model, the highest precision and specificity values were obtained in the RMS subclass, the highest sensitivity, F1-score, MCC and G-mean values were obtained from the EWS type tumor subclass (Table 4).

Table 4. ICA+ K-NN Performance metric values calculated from created models in the testing stage

Srbct tumor / Metrics	Ewing sarkomu (EWS)	Burkitt lenfoma (BL)	Nöroblastm (NB)	Rabdomiy osarkom (RMS)
Accuracy	0.9189	0.9067	0.8831	0.8947
Precision	0.8621	0.6364	0.7778	0.8800
Sensitivity	0.9259	0.7000	0.7368	0.8148
Specificity	0.9149	0.9385	0.9310	0.9388
F1-score	0.8929	0.6667	0.7568	0.8462
Matthew's correlation coefficient (MCC)	0.8291	0.6135	0.6803	0.7676
G-mean	0.9204	0.8105	0.8283	0.8746

Accuracy values obtained from the dimension reduction model made with ICA+ Forward Selection on K-NN model; for EWS, BL, NB, and RMS type tumors with 93.51%, 91.14%, 92.31% and 94.74% respectively. In the model created with ICA + Forward Selection + K-NN, the highest precision value was obtained in the EWS type tumor subclass, the highest specificity value in the BL type tumor subclass, and the highest sensitivity,

F1-score, MCC and G-mean values were obtained from the RMS type tumor subclass (Table 5).

Table 5. ICA+ Forward Selection+ K-NN Performance metric values calculated from created models in the testing stage

Srbct tumor / Metrics	Ewing sarkomu (EWS)	Burkitt lenfoma (BL)	Nöroblastom (NB)	Rabdomiyosarkom (RMS)
Accuracy	0.9351	0.9114	0.9231	0.9474
Precision	0.8966	0.8182	0.8333	0.8800
Sensitivity	0.9286	0.6429	0.8333	0.9565
Specificity	0.9388	0.9692	0.9500	0.9434
F1-score	0.9123	0.7200	0.8333	0.9167
Matthews correlation coefficient (MCC)	0.8611	0.6751	0.7833	0.8799
G-mean	0.9337	0.7894	0.8898	0.9499

Discussion

Although rare in childhood, cancer is still a major cause of death in children. In developed countries, only 0.5% of cancers occur in children under the age of 15 (16). Due to the long life expectancy in childhood and high treatment success rates in these cancers, cancers seen in childhood deserve special attention (17). Although the prognosis in childhood malignant soft tissue tumors mostly varies depending on the extent of the disease at the time of diagnosis, the region of origin of the tumor and the type of treatment chosen, the diagnosis and histological type of the tumor determine the patient's morbidity and mortality (18). Small round blue cell tumor, one of the childhood tumors, is a neoplasia with well-defined features in recent years. This malignant tumor, which shows a very aggressive course, is mostly observed in the adolescent age group and young adults (19). Despite aggressive multimodal treatment, median survival ranges from 17 to 25 months, with fewer than 20% of patients achieving 5-year survival (20).

Gene expression data obtained by microarray technology generally contain a large number of gene information belonging to a small number of patients. These data sets, which can be defined as high-dimensional for the methods used in data mining, reduce the model performance during the modeling phase. For this purpose, the performance of the classification models used is increased by obtaining genes with distinctive characteristics for the disease by performing dimension reduction analyzes before performing classification analyzes in gene expression data and the results obtained can be interpreted more easily (21).

In this study, the effects of PCA, ICA and PCA + Forward Selection, ICA + Forward Selection methods, which are among the dimension reduction methods on the open access gene expression data set of small round blue cell tumor types, on the K-NN classification method were examined and the results were compared.

In a study in the literature, after dimension reduction with PCA + Discrete Wavelet Transform (DWT) in the srbct gene expression data set, K-NN and Support Vector Machine (SVM) methods were used for classification. Accuracy rates were 92.73% for K-NN and 94.86% for SVM, respectively. When the results of this study in the literature are compared with the current study, it can be said that the model created with the proposed method, PCA + Forward Selection + K-NN method, is more successful in classifying the srbct gene expression data set (22).

When the results are compared, it can be said that the model established with PCA + K-NN has higher performance than the model established with ICA + K-NN. It is seen that the selection of variables with forward selection after PCA and ICA increases the model performance. The model that best predicts these four tumor types among all created models is the model established with PCA + Forward Selection + K-NN.

Conclusion

As a result, feature selection and feature extraction methods increase the prediction performance of machine learning methods by reducing the computational cost for gene expression data.

Ethics Committee Approval: Ethics committee approval is not required in this study.
Peer-review: Externally peer-reviewed.
Author Contributions:
Concept: F.H.Y., I.B.C., H.G.G.B., *Design:* F.H.Y., I.B.C., H.G.G.B., *Literature Search:* I.B.C., Z.T., *Data Collection and Processing:* I.B.C., Z.T., *Analysis or Interpretation:* F.H.Y., I.B.C., Z.T., *Writing:* F.H.Y., I.B.C., H.G.G.B., Z.T.
Conflict of Interest: The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.
Financial Disclosure: The authors received no financial support for the research and/or authorship of this article.

References

1. Gerald WL, Miller HK, Battifora H, Miettinen M, Silva EG, Rosai J. Intra-abdominal desmoplastic small round-cell tumor. Report of 19 cases of a distinctive type of high-grade polyphenotypic malignancy affecting young individuals. The American journal of surgical pathology. 1991;15(6):499-513.
2. Ordóñez NG. Desmoplastic small round cell tumor: II: an ultrastructural and immunohistochemical study with emphasis on new immunohistochemical markers. The American journal of surgical pathology. 1998;22(11):1314-27.
3. Bildirici K, Tel N, İhtiyar E, Algin C. Desmoplastic small round cell tumor. Journal of Cumhuriyet University Faculty of Medicine. 2002;24:87-90.
4. Amato RJ, Ellerhorst JA, Ayala AG. Intraabdominal desmoplastic small cell tumor: Report and discussion of five cases. Cancer: Interdisciplinary International Journal of the American Cancer Society. 1996;78(4):845-51.
5. Sequencing HG. Finishing the euchromatic sequence of the human genome. Nature. 2004;431:931-45.
6. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and intelligent laboratory systems. 1987;2(1-3):37-52.
7. Kocyigit Y, Korurek M. Classification of EMG signals using wavelet transform and fuzzy logic classifier. ITU Journal Series D: Engineering. 2005;4(3):25-31.
8. Blanchet FG, Legendre P, Borcard D. Forward selection of explanatory variables. Ecology. 2008;89(9):2623-32.
9. Han J, Kamber M, Pei J. Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems. 2011;5(4):83-124.
10. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine. 2001;7(6):673-9.
11. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Computers in biology and medicine. 2019;112:103375.
12. Song F, Guo Z, Mei D, editors. Feature selection using principal component analysis. 2010 international conference on system science, engineering design and manufacturing informatization; 2010: IEEE.
13. Guo Q, Wu W, Massart D, Boucon C, De Jong S. Feature selection in principal component analysis of analytical data. Chemometrics and Intelligent Laboratory Systems. 2002;61(1-2):123-32.
14. Bursa N, Tatlidil H. Evaluation of Independent Components Analysis from Statistical Perspective and Its Comparison with Principal Components Analysis. Journal of Suleyman Demirel University Institute of Science and Technology. 2020;24(2):474-86
15. Guo G, Wang H, Bell D, Bi Y, Greer K, editors. KNN model-based approach in classification. OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"; 2003: Springer.
16. Kutluk T. Epidemiology of childhood cancers. IU Cerrahpasa Faculty of Medicine Continuing Medical Education Activities, Diagnosis of Pediatric Cancers for All Symposium Series. 2006(49):11-5.
17. Picarsic J, Reyes-Múgica M. Phenotype and immunophenotype of the most common pediatric tumors. Applied immunohistochemistry & molecular morphology. 2015;23(5):313-26.
18. Dean A, Byrne A, Marinova M, Hayden I. Clinical outcomes of patients with rare and heavily pretreated solid tumors treated according to the results of tumor molecular profiling. BioMed research international. 2016;2016.

19. Tosun Yildirim H, Yildirim A, Diniz Unlu AG, Aktas S, Vergin C. Childhood Malignant Solid Soft Tissue Tumors; Diagnostic, Histopathological And Molecular Approach. Journal of Izmir Dr. Behçet Uz Children's Hospital. 2019;9(1):1-9.
20. Dufresne A, Cassier P, Couraud L, Marec-Bérard P, Meeus P, Alberti L, et al. Desmoplastic small round cell tumor: current management and recent findings. Sarcoma. 2012;2012.
21. Li G, Li J, Ju Z, Sun Y, Kong J. A novel feature extraction method for machine learning based on surface electromyography from healthy brain. Neural Computing and Applications. 2019;31(12):9013-22.
22. Nirmalakumari K, Rajaguru H, Rajkumar P, editors. Pca and dwt based gene selection technique for classification of microarray data. 2018 3rd International Conference on Communication and Electronics Systems (ICCES); 2018: IEEE.