


# Chapter 14

## Best Practices of Feature Selection in Multi-Omics Data

**Funda Ipekten**

 <https://orcid.org/0000-0002-6916-9563>

*Erciyes University, Turkey*

**Gözde Ertürk Zararsız**

*Erciyes University, Turkey*

**Halef Okan Doğan**

*Cumhuriyet University, Turkey*

**Vahap Eldem**

*Istanbul University, Turkey*

**Gökmen Zararsız**

*Erciyes University, Turkey*

### ABSTRACT

*With the recent advances in molecular biology techniques such as next-generation sequencing, mass-spectrometry, etc., a large omic data is produced. Using such data, the expression levels of thousands of molecular features (genes, proteins, metabolites, etc.) can be quantified and associated with diseases. The fact that multiple omics data contains different types of data and the number of analyzed variables increases the complexity of the models created with machine learning methods. In addition, due to many variables, the investigation of molecular variables associated with diseases is very costly. Therefore, selecting the informative and disease-related molecular features is applicable before model training and evaluation. This feature selection step is essential for obtaining accurate and generalizable models in minimum time with minimum cost. Some current methods used for feature selection are as follows: recursive feature elimination, information gain, minimum redundancy maximum relevance (mRMR), boruta, altmann, and lasso.*

DOI: 10.4018/979-8-3693-3026-5.ch014

## INTRODUCTION

Today, there is an increase in data in many areas. With this increase, the number and variety of the variables to be evaluated also increases. The increase in data and variables became a situation that needed to be solved among world problems. In addition, although there is a perception that having too much data in the scientific field, having too much information, correct information, or sufficient information may not be possible. However, it should not be forgotten that there is valuable information in a relatively large amount of data. It should be clear that it can be beneficial to have much data to extract this helpful information. However, performing data analyses to obtain and process this information can be difficult. In addition, its existence is a problem called the curse of data dimensionality (Verkeyesen M. and François D., 2005). High-dimensional data sets, where these problems are most common, are used successfully in multiple fields such as genetics, pharmacology, toxicology, nutrition, and genetics. The use of these high-dimensional data allows one to examine biology systems, cellular metabolism, and disease etiologies in more detail. However, the number of samples ( $n$ ) of these data is considerably lower than the number of variables ( $p$ ) and the heterogeneity of the data, the missing observations in the data as a result of the use of high-output technology, limits the use of traditional methods that can be used in this field. Therefore, there is a need for the clinical understanding of the biological system based on research and machine learning, and statistical learning methods to analyze this clinical information statistically (Hastie et al., 2009). Several studies are show that machine learning methods are used and applied successfully in studies carried out in this field. Some of these studies are listed in Table 1.

*Table 1. Some studies using feature selection*

Datasets		Methods	References	
Ovarian cancer	Classification	MKL	Wilson et al.	2019
Breast cancer	Classification	MKL	Tao et al.	2019
Gene expression	Classification	SVM	Golub et al.	1999
Gut microbiota	Classification	RF	Franzosa et al.	2019
Colon cancer	Classification	SVM	Moler et al.	2000
Ovarian, leukemia, colon	Classification	SVM	Furey et al.	2000

MKL: Multiple Kernel Learning, SVM: Support Vector Machine, RF:Random Forest

In general, in all of the studies given in Table 1, researchers aim to optimize the classification of disease-related samples, produce models that can be used to predict system behaviors, or properties or provide the most accurate result appropriate in terms of classification performances. However, the large number of variables in the data used can complicate the structure of the models to be created hence reducing their accuracy. In addition, because the number of variables is too large, the investigation of disease-related genes or other omics causes considerable losses in terms of both time and cost (Hastie et al., 2009). Therefore, it is not always possible for researchers to carry out these studies in depth. Therefore, feature selection is made before model training is evaluated to make the models obtained with learning methods more generalizable, predictable, and ineffective against noisy values, with minimum time and

14 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:  
[www.igi-global.com/chapter/best-practices-feature-selection-multi/342532](http://www.igi-global.com/chapter/best-practices-feature-selection-multi/342532)

## Related Content

---

### Clustering Genes Using Heterogeneous Data Sources

Erliang Zeng, Chengyong Yang, Tao Liand Giri Narasimhan (2010). *International Journal of Knowledge Discovery in Bioinformatics* (pp. 12-28).

[www.irma-international.org/article/clustering-genes-using-heterogeneous-data/45163](http://www.irma-international.org/article/clustering-genes-using-heterogeneous-data/45163)

### Machine Learning Applications for Classification Emergency and Non-Emergency Patients

Zeynel Abidin Çiland Abdullah Caliskan (2024). *Research Anthology on Bioinformatics, Genomics, and Computational Biology* (pp. 1090-1103).

[www.irma-international.org/chapter/machine-learning-applications-classification-emergency/342564](http://www.irma-international.org/chapter/machine-learning-applications-classification-emergency/342564)

### Multidimensional Numbers and the Genomatrices of Hydrogen Bonds

Sergey Petoukhovand Matthew He (2010). *Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications* (pp. 193-206).

[www.irma-international.org/chapter/multidimensional-numbers-genomatrices-hydrogen-bonds/37902](http://www.irma-international.org/chapter/multidimensional-numbers-genomatrices-hydrogen-bonds/37902)

### Levy Processes Simulation for Modeling in Bioinformatics

Anna V. Kuzmina (2020). *International Journal of Applied Research in Bioinformatics* (pp. 55-64).

[www.irma-international.org/article/levy-processes-simulation-for-modeling-in-bioinformatics/260827](http://www.irma-international.org/article/levy-processes-simulation-for-modeling-in-bioinformatics/260827)

### Molecular Docking Study of Expansin Proteins in Fibers of Medicinal Plants Calotropis Procera

Anamika Basu (2020). *International Journal of Applied Research in Bioinformatics* (pp. 10-17).

[www.irma-international.org/article/molecular-docking-study-of-expansin-proteins-in-fibers-of-medicinal-plants-calotropis-procera/261866](http://www.irma-international.org/article/molecular-docking-study-of-expansin-proteins-in-fibers-of-medicinal-plants-calotropis-procera/261866)