EXERCISE SCIENCE 운동과학
Vol.30, No. 2, May 2021: 131-146
https://doi.org/10.15857/ksep.2021.30.2.131
ISSN(Print) : 1226-1726
ISSN(Online) : 2384-0544
REVIEW ARTICLE

Machine Learning Application in Genomic, Exercise, and Vital Datasets

Kyung-Wan Baek PhD[1,2], Jung-Jun Park PhD[1], Jeong-An Gim PhD[3]

[1]Division of Sport Science, Pusan National University, Busan; [2]Department of Physical Education, Gyeongsang National University, Jinju; [3]Medical Science Research Center, College of Medicine, Korea University, Seoul, Korea

PURPOSE: Machine learning (ML) refers to newly developed computer algorithms that are improved through iterative experiences. ML applications are expected to assist humans in analyzing large amounts of data. This review has outlined the application of ML in analyzing variable vital data such as walking steps, exercise intensity, heart rate, sleeping hours, sleep quality, resting heart rate, blood pressure, and calorie consumption in a day. Vital data consist of different variables that are closely related to genomic or exercise data. The prediction of healthy traits from a vital dataset has become a necessity in personalized medicine.

METHODS: Considerations and repeated tasks in supervised, semi-supervised, and unsupervised ML methods are presented. ML methods such as artificial neural networks, Bayesian networks, support vector machines, and decision trees have been widely used in biomedical studies to develop predictive models. Through vital data, these models can help in effective and accurate decision-making for a healthier life.

PURPOSE: Models based on genomic, exercise, and vital datasets provide a healthy lifestyle through regular exercise. We have provided guidelines to help in the selection of these ML methods and their practical application for variable vital data analysis.

CONCLUSIONS: Our guidelines could serve as a foundation for implementing both participatory medicine and data-driven exercise science.

Key words: Machine learning, Genomics, Vital, Datasets, Exercise science

INTRODUCTION

Data processing speed is increasing as computer performance improves, and big data is accumulating as storage space is expanded. With the development of cloud computing and wireless communication networks, it is now possible to collect data from a variety of sources [1]. Machine learning (ML) libraries are being developed to provide accurate insights obtained from various types of big data [2]. These data contain various types of vital data, and it is now time to discuss how to use the collected vital data [3,4]. In this paper, we discuss how these data can be used in participatory medicine and what methods are available. Currently, there is a favorable environment for acquiring and utilizing vital big data. Wearable devices and open genomic databases can easily and quickly collect and use vital big data.

Wearable devices are electronic devices with microcontrollers that can be implanted into clothing or worn as accessories on the human body. Wearable devices have sensors that detect human vital data and then send it to other devices. To date, a large amount of vital data has been gathered [5,6].

Genomic information can be generated using next-generation sequencing (NGS) methods. Various NGS machines, such as whole-genome sequencing (WGS), RNA-sequencing (RNA-seq), and Methyl-se-

Corresponding author: Jung-Jun Park **Tel** +82-51-510-2713 **Fax** +82-51-510-3747 **E-mail** jjparkpnu@pusan.ac.kr
Corresponding author: Jeong-An Gim **Tel** +82-2-2626-2362 **Fax** +82-2-2626-1962 **E-mail** vitastar@korea.ac.kr
*This work was supported by a 2-year research grant from Pusan National University.
Received 31 Mar 2021 **Revised** 14 May 2021 **Accepted** 17 May 2021

quencing (Methyl-seq), can generate large amounts of information in a short period of time [7]. Microarray can also be used to analyze genotyping, gene expression, or DNA methylation [8]. The WGS and genotyping results were presented as nominal variables, whereas gene expression and DNA methylation levels through RNA-seq, methyl-seq, and microarrays were represented as continuous variables. Genomic data, such as continuous variables or nominal variables, should be thoroughly understood because proper libraries are required for analysis based on the type of genomic information. Genomic data are used as input for ML and aids in phenotype classification, regression, and clustering [9,10].

Recently, such genomic data has been deposited in public databases. Public databases, such as TCGA, contain oncogene data [10]. NCBI GEO (https://www.ncbi.nlm.nih.gov.geo/) and ArrayExpress (https://www.ebai.ac.uk/arrayexpress/) have accumulated data under various sample conditions [11,12]. To date, genome data analysis results have been uploaded to various disease samples, such as cancer. Post-exercise comparative analysis data were also uploaded [13,14]. Scientists have paid close attention to research that analyzes and accumulates genomic data quickly and accurately, whereas research that seeks insights from the accumulated data has not.

To date, many vital data have been accumulated. ML technologies are used to provide insights from these datasets. The data sources and types used in this study are as follows.

- NGS data can be used to obtain genome sequencing data, gene expression, and DNA methylation patterns of whole genomic regions using WGS, RNA-seq, and Methyl-seq. WGS data provide the genotype, which is considered a categorical variable. Gene expression and beta values (methylation level of CpG islands) were obtained using RNA-seq and Methyl-seq. These two variables were viewed as continuous variables.
- Microarray data can be used to determine the genotype, gene expression level, and methylation level of specific genomic loci using DNA chips, gene expression chips, and methylation chips, respectively. DNA chips provide genotypes as categorical variables, whereas gene expression and methylation chips provide continuous variables.
- Vital data refers to all raw facts derived from living organisms, especially all figures necessary for survival and health maintenance.
- Wearable devices track exercise and vital data such as the number of steps taken while walking, exercise intensity, heart rate, sleeping hours, sleep quality, resting heart rate, blood pressure, and calorie consumption in a day.
- Electronic health records are vital data for patients and populations that are systematically collected and stored in digital form electronically. These records can be shared by different healthcare platforms. Data can be shared in the network and management systems.
- Questionnaires are used to collect data about participants' lifestyles, such as physical activity, sleep habits, chronotype, and diet.

Our first aim was to obtain insights from genomic, exercise, and vital data using ML. The second aim was to teach participants how to live healthy lifestyles so that personalized and participatory medicine can be realized. Thus, the accurate prediction of a health state and maintenance of a healthy state is an intriguing and challenging task for scientists. ML technologies have aided in the realization of personalized and participatory medicine. To maintain a healthy state, a proper exercise state can be treated by appropriate status recognition. In this paper, we discussed studies on disease prediction, prognosis, and applications to exercise-related traits. We expected that the technologies described in this systematic review will assist in the discovery and identification of patterns in exercise-related datasets.

Machine Learning Technologies

ML is an artificial intelligence (AI) technique. ML is used to learn and find patterns in large datasets and provide insights to humans [15, 16]. ML has been used in biomedical studies for many applications, such as disease prognosis or diagnosis, biological sample clustering, and pattern detection in medical images [17-19]. ML can be applied to a wide range of research areas by employing a variety of techniques and algorithms. ML methods can be divided into two main types: supervised learning and unsupervised learning. Supervised learning makes use of a labeled set of training data, whereas unsupervised learning makes no use of a labeled dataset. Therefore, the main issues of supervised learning are classification and regression [20-22], which is the main theme of unsupervised learning and provides similar features that they share from a complex dataset [23,24]. In this section, we discuss ML technologies used for genomic, exercise, and vital data.

1. Classifier model based on ML

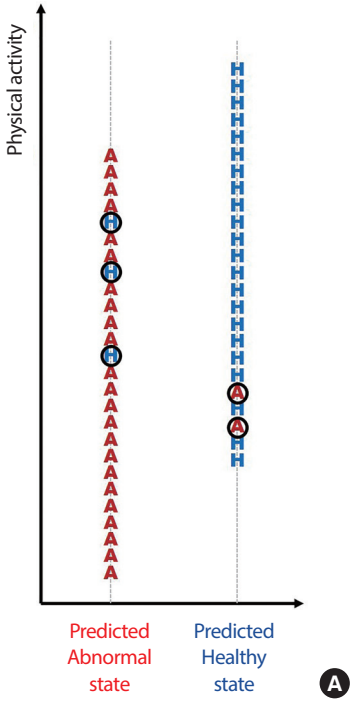
The main aim of ML application is to design a model capable of performing classification, prediction, pattern recognition, and estimation of nominal or continuous variables. Classification is a commonly used task in ML applications [20-22]. Classifiers learn how to classify features into

two or more predefined classes through learning processes. When the classification model is designed, training errors can occur.

We assume a simple model that predicts a healthy state based on physical activity (Fig. 1). Using vital data collected from wearable devices, we could measure physical activity, which can be applied to features as a predictor of health status. We also collected genomic data from public databases, such as genotype, gene expression, and DNA methylation. In the model, the input data can be nominal, continuous, or both variables, whereas the output will be a nominal variable (abnormal state or health state). Fig. 1 illustrates the classification process of an abnormal or healthy state. The circled subjects represent any misclassification of the type of health (or abnormal health) state by the ML model.

The ML model differentiated between abnormal and health states, and one continuous variable (physical activity) with multiple features was identified (Fig. 1A). According to the ML model, more physically active participants tended to be healthier. Typically, the predicted and diagnosed states can be presented as a confusion matrix, which provides true and false positives and negatives, respectively (Fig. 1). The model's performance can be estimated as sensitivity (recall), specificity, precision, accuracy, or F1 score, where the F1 score is a representative value of the performance of the model's overall performance. As discussed, the performance of the classification model is estimated in terms of sensitivity, specificity, and accuracy. Generally, sensitivity is defined as the ratio of true positive to true, whereas specificity is defined as the ratio of true negative to actually false (Fig. 1B). These terms can be represented graphically as a receiver operating characteristic (ROC) curve and evaluated as the area under the curve (AUC). AUC is the integral of the ROC curve and is used to evaluate the overall performance of the classification model. Accuracy is proportional to the total number of correct predictions, and AUC is the model's performance based on the ROC curve, which plots the point of contact between sensitivity and 100-specificity (Fig. 2). To visualize changes in specificity and sensitivity, ROC curve provides two performance values of the model evaluation. The model performs better when the curve is closer to the upper left corner. In other words, a good model has high sensitivity and specificity. It should also be located at the top left of the y=x graph for best performance.

We have discussed how to design and evaluate a model that can accurately predict the health state based on vital data. The data to be obtained in the field of exercise science will be accumulated and collected. Developing a model that can monitor optimal health status and recom-



	Diagnosed Abnormal state	Diagnosed Healthy state
Predicted Abnormal state	TP, 23	FP, 3
Predicted Healthy state	FN, 2	TN, 22

$$Sensitivity(Recall) = \frac{TP}{TP+FN} = \frac{23}{23+2} = 92.0\%$$

$$Specificity = \frac{TN}{TN+FP} = \frac{22}{22+3} = 88.0\%$$

$$Precision = \frac{TP}{TP+FP} = \frac{23}{23+3} = 88.5\%$$

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} = \frac{23+22}{23+2+3+22} = 90.0\%$$

$$F1\ score = 2 * \frac{1}{\frac{1}{Precision}+\frac{1}{Recall}} = \frac{1}{\frac{1}{0.885}+\frac{1}{0.92}} = 0.902$$

B

Fig. 1. (A) Classification of abnormal and health state based on the machine learning model. The X-axis is healthy, and the Y-axis is physical activity, one of the features as continuous variable used in the model. (B) Confusion matrix of the model. The sensitivity is the ratio of actual abnormal state to predictive abnormal state and known as recall. The specificity is the ratio of actual health state to predictive health state. Precision is the ratio of the model that classifies abnormal state from what is actually abnormal. Accuracy represents the ratio of the correctly classified model. The F1 score is harmonic mean of sensitivity and precision, which gives a numerical indication of the overall performance of the model.

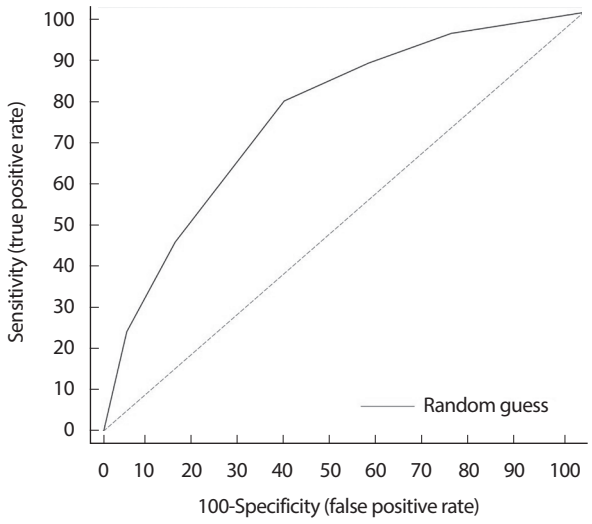


Fig. 2. Changes in specificity and sensitivity—two performance value of model evaluation—were visualized as receiver operating characteristic (ROC) curve. In this figure, the overall performance of the model is evaluated as the area under the curve (AUC). The ROC curve should be at the top left of the y=x graph and the AUC should be at least 0.5 to ensure performance as a classification model. The x-axis is 100-specificity (false positive rate), and the Y-axis is sensitivity (true positive rate).

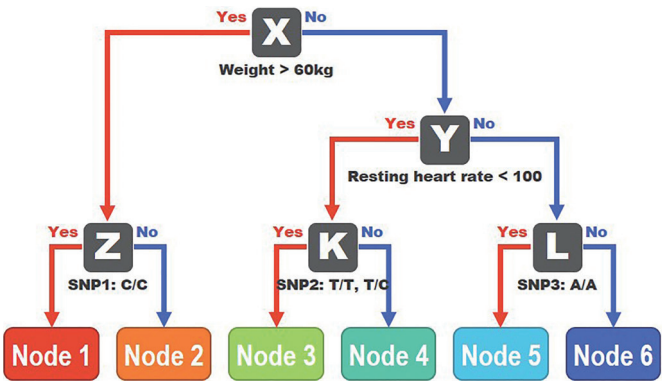


Fig. 3. A decision tree. One root node is located at the highest part and is indicated as "X" in the round square. Four intermediate nodes make clades in the tree. Each node can classify both nominal and continuous variables. In the nodes, X and Y nodes classify continuous variables, whereas Z, K, and L nodes classify nominal variables. Terminal nodes are indicated as numbers in the round rectangle. Total data number, such as total number of participants in the terminal node is the same as root node. In this example, the six terminal nodes mean the six subsets of total data.

mend appropriate exercise habits based on such large amounts of data will be aid in the development of the model in this review.

2. ML methods

1) Decision tree

A decision tree (DT) is a tree-structured classification pattern in which nodes represent the input variables and the leaves represent decision outputs. DT is a prominent ML method that has been widely used for classification. DT is easy to interpret and learn, and many analysis libraries are now available [25,26].

DTs have two advantages. First, DTs visually and intuitively classify the data. Second, DTs are applicable for both classification and regression. In other words, categorical or continuous values were classified.

After confirming that the new data belongs to a specific terminal node, the new data are classified into the most frequent category in the corresponding terminal node in the DT's category prediction process. As shown in Fig. 3, the terminal nodes are determined by whether they have a specific genotype of single-nucleotide polymorphisms 1, 2, and 3 (intermediate nodes Z, K, and L, respectively). Similarly, in the case of regression, the data are classified by using the mean of the dependent variable at each node as a predicted value. Data are sorted through root node X, which classifies participants whose weight is greater than 60 kg, and intermediate node Y, which classifies participants whose resting heart rate is less than 100.

DTs are more likely to work well with specific data because the decision boundary is perpendicular to the data axis. Random forest, a technique for creating multiple DTs for the same data, synthesizing the results, and improving the prediction performance, emerged as a solution to this problem [27,28]. Whether it is a DT or a random forest, major languages such as R and Python are supported in the form of libraries, and they support the visualization of quality that can be used in papers or research reports.

2) Artificial neural networks

Artificial neural networks (ANNs) can be used to classify or recognize patterns in nominal or image data [29,30]. Artificial neural connections are provided by multiple hidden layers. Fig. 4 presents the structure of an ANN with a network of interconnected nodes. Each node is connected to neurons, each of which has its own weight. Hidden layers receive input data and provide output in a black box. Each neural network is defined as a matrix, which is multiplied by the input matrices to produce the output. The layer containing the input vector is known as the input layer, the layer containing the final output value is known as the output layer, and all layers located between the input and output layers are known as hidden layers. In Fig. 4, four layers are shown, but the input layer is not included when counting the number of ANN layers; therefore, there are three layers in the example. ANNs with two or more layers are referred to as "MLP (multi-layer perceptron)". As the number of

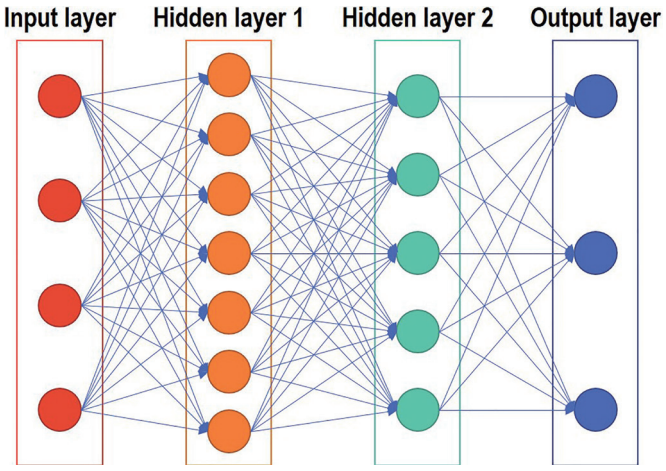


Fig. 4. Artificial neural networks. The feature enters the node located in the input layer, then the neural network produces the output through the calculations. The feature as input data should be expressed in the form of a matrix, and each calculation can be expressed in the form of matrix multiplication. In this model, there are four nodes in the input layer, seven and five nodes in the two hidden layers, and three nodes in the output layer. Arrows represent artificial neurons, which perform calculations on the input data and export them to the output data. The circle represents a node and applies the activation function to the data received from the previous artificial neuron and delivers it to the next artificial neuron.

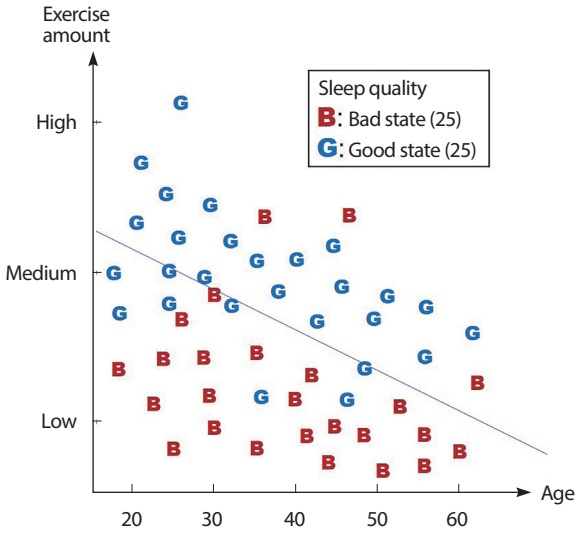


Fig. 5. The support vector machine algorithm is a binary linear classification model that determines which category a new data belongs to base on a given dataset. The example presented in this figure is a model that divides sleep quality according to age and exercise intensity. The good and bad states are given for the training data, and the optimal hyperplane is obtained after measuring the distance between the two data in the two groups to find the center of the two data. As in this example, a linear classification model is applied if it can be divided by a straight line, and a nonlinear classification model is used when it cannot be divided by a straight line.

hidden layers increases, the ANN is called the "deep" state, and the ML paradigm that uses this deep ANN as a learning model is called "deep learning" In addition, a deep enough neural network used for deep learning is referred to as "deep neural network (DNN)" [31].

DNNs have been recognized as powerful classifiers in several application areas, including genomic information, and they are used to classify normal and diseases-related features from large and complex genomic data. The input is in form of vectors containing genomic information from the individual's selected genomic regions, such as genotypes. To date, some tools for detecting disease states from genomic data or images. DeepSEA, a multitask hierarchically structured CNN trained on large-scale functional genomics data, was able to learn sequence dependencies at multiple scales and simultaneously produce predictions of DNase hypersensitive sites, transcription factor binding sites, histone marks, and the influence of genetic variation on these regulatory elements, with a level of accuracy superior to those of other tools for prioritizing non-coding functional variants [32]. DeepVariant, a CNN-based variant caller trained directly on read alignments without any prior knowledge of genomics or sequencing platforms, recently outperformed standard tools on a variety of calling tasks [33]. SpliceAI, a 32-layer DNN, can predict both canonical and non-canonical splicing directly from exon–intron junction sequence data [34]. Although the classification accuracies of these tools are insufficient to drive clinical reporting, they help guide the interpretation of clinical genomic data by prioritizing potential candidate variants for further consideration.

3) Support vector machines

Support vector machines (SVMs) are relatively new techniques that have been used in cancer prediction and prognosis [35,36]. SVMs can provide separate hyperplanes. Fig. 5 depicts the classification of good and bad states based on age and amount (or intensity) of exercise. The identified hyperplane can provide a choice between two distinct features. SVM has a high classification accuracy and suggests a visual classification method for planes (or space). Since the early 2,000 s, when high-throughput microarray gene expression data became available, SVM has been used as a classifier in cancer classification. The first attempt used a linear SVM to classify two different types of leukemia using gene expression microarray data [37]. In this study, 38 patients' data were used as the training set, and the learning algorithm was trained to distinguish between the two labeled forms of leukemia. The Affymetrix Hgu 6,800 chips covered 7,129 gene expression probes and were used as features. Each feature was assigned to a weight based on its relevance to the two

classes. The learned SVM model was used as the test set, and another independent dataset of 34 patients was used as the control set. The performance of SVM in classifying high-dimensional (gene features) and small sample size data was first tested in this study. Although this study was attempted in the last two decades, the methods of this study have yet to be attempted. The SVM was applied to colon cancer tissue to classify the selected features. They used a collection of 40 colon cancer tumors and 22 normal colon tissues [38].

4) Principal component analysis

It is common to have many dimensions (variables as columns; participants as rows) in genomic, exercise, and vital data. Genomic data is divided into two parts: a genetic part (e.g., genotype, gene expression level, DNA methylation level, etc.) and participant information (e.g., blood pressure, body mass index, whole blood counts, etc.). As a genetic component, RNA-seq or microarray data have at least 10,000 columns. Thus, selecting the overall gene expression patterns is difficult. Principal component analysis (PCA) allows for the simultaneous plotting of all dimensions and the detection of patterns [39,40] (Fig. 6).

When developing a model, such as regression or DT, multicollinearity with high correlations between variables will result in the incorrect model, causing problems with interpretation [41]. If multicollinearity exists, PCA can be used in a model developmet to reduce highly correlated variables into a single principal component. Therefore, it is important to screen the variables for highly correlated variables, such as body weight and blood lipid profile, height and body weight, and total cholesterol and triglyceride levels.

PCA can be used to predict the onset of a fatal state by analyzing the change in distribution or trend in time series vital data. For example, it is used to detect anomalies in vital data results such as blood glucose levels, blood pressure, and whole blood cell analysis.

PCA is an unsupervised learning technique and dimension-reduction method. PCA enables the visualization ofclassification data as well as the analysis of phenotypes. PCA decomposes the total genomic information (such as genotype) into N axes of genomic information known as principal components. The principal components of population genomics can correspond to evolutionary processes such as population divergence [42,43].

3. Pre-processing, algorithm selection, and feature treatment of genomic, exercise, and vital data

Raw genomic, exercise, and vital data from open genomic databases, electronic health records, and wearable devices are required to improve data quality. To use ML, two processes must be carried out: pre-processing and algorithm selection. First, the data must be provided in matrix form. A row in the matrix can represent each individual's ID, and a column can represent the characteristics of each variable (or vice versa). Missing values are then treated as deletions or imputations. The scientists must then decide on the algorithm's input data. In this step, vital data can be converted into a standardized and typical format. Dimensionality reduction, feature selection, and feature extraction are three important approaches. Additionally, we discuss whether each variable follows a normal distribution, distribution patterns, and distribution

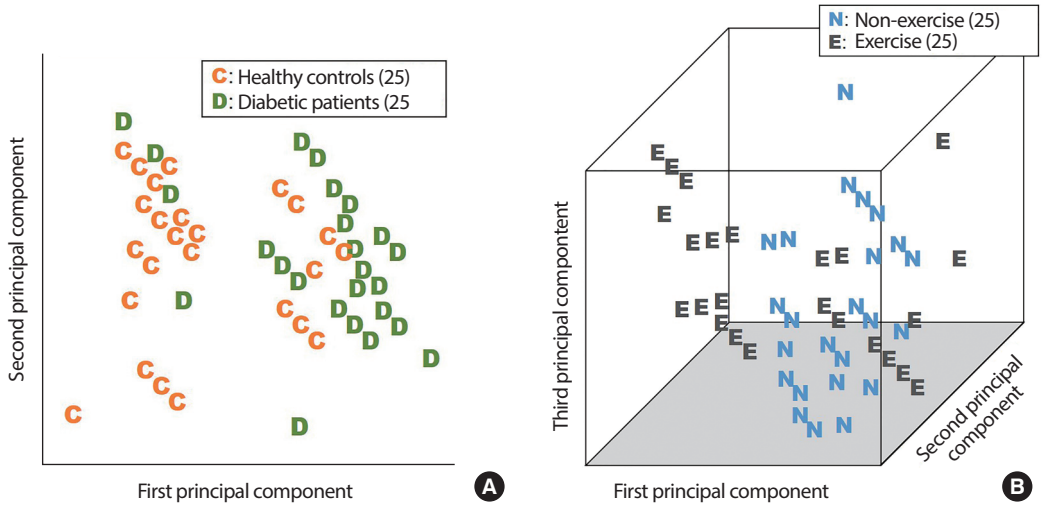


Fig. 6. (A) Two-dimension principal component analysis (PCA) (2D PCA), and (B) three-dimension PCA (3D PCA). Based on the genomic data, the discovered features (the number of each gene) are reduced, and the two groups are distinguished from two (A) to three (B) components.

pattern visualization.

1) From vital data to features

Health logs, exercise, vital signs, and gene expression data are generally classified as categorical or continuous variables. A matrix of rows and columns can be used to represent a collection of data, where a row can represent each individual's ID and a column can represent the characteristics of each variable (or vice versa). First, we determine whether each piece of data is a matrix. All data must be entered in matrix form, whether manually, through a specially designed input system, or automatically from a variety of samples such as microarrays. Second, we must determine whether each variable is a categorical variable or nominal variable. This is because the analysis strategy is determined by whether the input and output variables are categorical or nominal. This means that ML libraries receive nominal and quantitative data as input and output categorical or nominal variables. Genotype data were classified as nominal data, and gene expression or DNA methylation data were classified as quantitative data. Disease or health state was classified as nominal data, and proper exercise level was provided as quantitative data. Finally, proper selection of ML libraries was also important. Many ML libraries based on Python or R are available, and users should be familiar with the input and output as well as the algorithms of each library.

ML approaches may be effective in the analysis of large and complex datasets [44,45] and are likely to become important to genomics as larger datasets become available through large genome projects, such as the 1,000 genome projects, and 100,000 genome projects in the United Kingdom. Epigenomic data are also available for the ENCODE project. However, these datasets were provided as sequencing data and the participant's vital data. Sequencing data can be converted into genotype, gene expression, and DNA methylation levels. These processes are beyond the scope of this paper, but they are covered in numerous review papers and handbooks [46,47].

2) Handling missing data

Missing values can be collected from the genomic, exercise, and vital data. The DNA chip collects genomic data, which is not referred to as genotypes. The detection limits of the instruments result in missing data. There are two approaches to dealing with missing values: deletion and imputation. To remove missing values, all rows or columns were deleted (Fig. 7). If a participant has a missing value, you can either delete all

ID_REF	GSM14XXX34	GSM14XXX35	GSM14XXX36	GSM14XXX37	GSM14XXX38	GSM14XXX39	GSM14XXX40	GSM14XXX41	GSM14XXX42	GSM14XXX43	GSM14XXX44	GSM14XXX45
Age	67	77	74	73	80	68	83	64	87	87	83	81
Gender	Female	Male	NA	Male	Male	Female	Male	Female	Male	Male	Male	Female
Income	16.29	14.64	NA	15.28	15.3	15.93	15.02	16.93	NA	16.1	14.22	13.18
Smoke	N	Y	NA	N	Y	Y	Y	N	N	N	N	N
Alcohol	N	Y	Y	Y	Y	Y	NA	N	Y	Y	Y	Y
BMI	23.2	30	22.6	28.1	41.1	36.8	23.4	25.8	26.6	NA	36	29.9
Diet	HP	MP	MP	MP	MP	LP	LP	MP	MP	MP	MP	MP
ILMN_XXX345XXX	14.95632	15.01515	14.78805	15.10025	14.94286	15.04881	15.04881	14.9313	15.04881	7.249382	7.037813	7.135032
ILMN_XXX346XXX	10.89202	11.36086	11.01068	11.37837	10.70285	11.45785	11.22959	10.91332	11.13302	7.370524	7.296496	7.750929
ILMN_XXX347XXX	7.029775	NA	7.082129	7.171403	7.156152	7.475266	7.056772	7.417414	7.126659	7.284088	7.331829	6.547495
ILMN_XXX348XXX	7.067703	NA	7.100199	7.132482	NA	7.385033	NA	7.157782	NA	10.11846	NA	NA
ILMN_XXX349XXX	7.511303	NA	7.273023	7.256698	7.320747	7.431908	7.228571	7.32484	7.621908	6.512435	NA	7.266859
ILMN_XXX350XXX	7.743872	7.492396	7.446314	7.635925	7.307447	7.397352	7.547048	7.439077	7.445051	8.636507	8.586738	7.106121
ILMN_XXX351XXX	13.39975	13.58255	13.78422	13.60584	13.44975	13.5486	13.5527	13.37828	13.41576	7.346921	7.317947	6.555672
ILMN_XXX352XXX	8.728905	8.91229	8.600676	8.834941	8.707664	8.531401	8.993117	8.755519	8.395389	6.474998	NA	7.143695
ILMN_XXX353XXX	7.245316	7.273272	7.164616	6.969526	NA	7.083971	7.048417	7.25025	6.760887	8.835571	NA	6.654364
ILMN_XXX354XXX	7.395432	7.564133	7.058695	7.250724	7.295159	7.484064	7.418875	7.398098	7.482954	9.760397	NA	8.814406

Fig. 7. An example of vital data (based on electronic health records or questionnaires) and genomic data. Sample information and genome information organized as a table. In the table, the column name is the sample ID, and the row name is the vital data and genome data. It is better to code the sample ID as alphabet letters+number format according to the same number of digits. Sample ID should contain the characteristics of the sample to characterize sample traits. The vital data is divided into continuous variables (age, height, body weight, body mass index, etc.) and categorical variables (gender, smoking, alcohol consumption, etc.). Genomic data is also divided into continuous variables (gene expression level resulting from RNA-seq, beta value from methylation analysis, etc.) or categorical variables (such as DNA chip or genotype resulting from targeted sequencing). In the example, some of the microarray analysis results that can be interpreted as gene expression levels are presented. In the case of the gene expression level, the variation may be large for each value, but it can be reduced by processing the log. Some missing values in the vital data and genome data could be detected, and they were marked as "NA". In the case of samples with many missing values, it is possible to drop the samples itself, such as a column marked with a red shade. If there are many missing values in the sample data or genome data, the sample data or the specific gene itself can be dropped like a row marked in blue shade. For the remaining samples, it can be replaced with an average value or zero, depending on the characteristics of the data.

the participant's rows or delete the variable with the missing value. This method can be used when there are many participants or variables, however, when the sample size or number of variables is small, the power of the model is reduced. If a missing value occurs in a categorical variable, the categorical frequency may be replaced with the mean, median, or maximum value, but a bias may occur if the ratio of missing values is high. For example, if the ratio of women to men is high, replacing missing values with women may result in greater bias. If a missing value occurs in a continuous variable, it can generally be replaced by the median or mean. Furthermore, we can predict the value in a missing column by analyzing the pattern in that column. If the proportion of missing values in a single variable is low, it can be imputed to the mean, median, or zero. If the number of missing values is high, the corresponding column (or row) should be deleted. It is important to know how to handle missing values, and it is always possible that imputation can be replaced by an incorrect value [48].

All strategies for processing vital and genomic data are depicted in Fig. 7. Simplified data, processing of missing values, and approximate data types were provided.

3) Feature selection: application of personal information as features

Individual characteristics such as height, body weight, blood pressure, and disease status, can be correlated with results such as quantified gene expression, exercise capacity, and health status. The result can be expressed using correlation analysis, with the former as the independent variable and the latter as the dependent variable. In this case, the parameters used in the analysis are called features, and it is important to select an appropriate feature for ML analysis [49,50].

The user should choose the appropriate data for the ML library as input and continue to optimize the appropriateness of the output data. ML can be thought of as a black box that connects input and output [51]. This black box is a function of the input data and can be linear or nonlinear in nature. We use training data to learn this function, but it is not always well learned. For example, we might want to build an ML model that recommends an appropriate exercise regimen basing on each individual's height as input data. Is it possible to accurately predict what exercise and how much exercise is done based solely on height? Perhaps each individual's height is insufficient to provide appropriate exercise strategies. In other words, if the user does not continue to consider the appropriateness of the input and output, incorrect data will be learned, and incorrect results will emerge. In the previous example, we discussed that the performance of ML is highly dependent on the amount and quality of the data. The best input data contains only adequate information that is neither too short nor too long, and some degree of correlation between the input and the result should be predicted. If this is the case, can we only obtain adequate and accurate information through observation. Of course, if we have complete prior knowledge (or background) of the problem we are attempting to solve, we can only gather the right information; however, in most cases, this is not the case, and health log data is no exception. Perhaps in many cases, there is insufficient background knowledge on the problem to be solved, and the application of ML techniques is intended to compensate for the incompleteness of the instrumentation.

Thus, one solution is to collect enough previous data from various sources. For example, to measure the health log, we used a sufficient number of sensors. After collecting the data, we determine whether the features are useful. The process of determining whether a feature is useful is known as feature selection or feature extraction. Because this process generates new input data based on existing input, it is usually run prior to the learning process and requires refined data. Therefore, this process is critical to the machine learning architecture. Next, we select the input and output variables, which we cover in the next section.

Machine learning technologies perform better when the dimensionality is lower. In addition, dimensionality reduction can erase irrelevant features, lower noise, and produce more accurate learning models because of the involvement of fewer features. PCA and multidimensional scaling can solve the dimensionality reduction problem by using vital data from higher to lower dimensions. Feature selection is a method of reducing dimensionality by selecting new features that are a subset of the old ones. Embedded, filter, and wrapper approaches can be used as the three main methods of feature selection. A new set of features can be created from the initial set that captures all the significant information in a dataset using feature extraction. New sets of features can be created and then made available to gain the described benefits of dimensionality reduction.

Because vital datasets will be complex and diverse in the near future, feature selection will be a critical process. High-dimensional datasets are plagued by the curse of dimensionality, which involves overfitting. The overfitting model outperforms the training data, but it produces fewer generalized patterns.

4) Provide personalized lifestyle solution

Our aim is to develop a model that provides a suitable exercise program using genomic and vital big datasets. As previously explained, the correlation between the input and output must be predicted, and sufficient input data are required. Recent public databases, such as NCBI GEO, have bio signal-related variables such as lifestyle, such as age, smoking, drinking, socioeconomic status, body mass index, disease status, and other variables such as genotype, gene expression, and DNA methylation. Many vital datasets have been developed. In addition, data such as changes in gene expression based on individual exercise amount, exercise type, and exercise program have been accumulated in previous studies [52,53]. A ML model that predicts the intensity of individual exercise based on gene expression data collected before and after the exercise program for a specific period may be proposed. In the future, it is expected to build a model based on various long-term health log data that suggests appropriate behaviors, eating habits, and exercise habits by learning factors related to health status. A model that detects the abnormal health states could be developed, and used to provide the time to go to the hospital and the emergency state of elderly people [54].

This section presents a methodology for applying ML to various data sources. Using this, we can transform the data into a format that the machine can recognize and suggest a way to extract insights from the data. We can extract insights that humans cannot recognize from data that simply lists individual characteristics by applying this methodology to various data. It can help to advance participatory medicine by recommending a lifestyle or exercise plan that is tailored to each individual.

5) Subconclusions

To date, many ML technologies and algorithms have been developed for disease prediction, diagnosis, and prognosis. ML technology has been used to identify a larger number of biomarkers [49,50]. In the field of exercise science, input data will continue to accumulate, and data interpretation and application using machine learning will become faster and faster.

4. Examples of application to exercise science using NCBI GEO datasets

The NCBI GEO database contains RNA-seq and microarray analyses of cancer and other disease samples. Omics data accumulated in many diseases or cancers have been deposited, and the results of the correlation between the dataset and genome collected through wearable devices are expected to be uploaded. In this review, we used the datasets in the GSE8479 study, and confirmed the change in human skeletal muscle after exercise training of older (OE) samples [52]. In the GSE28422 study, the effect of resistance exercise and resistance training on the skeletal muscle transcriptome of young and old participants was detected [53]. The authors of the two previous studies examined gene expression patterns in accordance with the study objectives, and we reanalyzed gene expression patterns using machine learning techniques. R studio was used for all analyses. In this section, we will explain the re-analysis method using two publicly available omics datasets, the language and library used, and the appropriate visualization and key factor discovery strategies based on the data.

So far, NCBI GEO provides the results of omics analysis on samples of various traits, and you can download and use data related to the subject of interest by searching for it with an appropriate keyword. It can be used as prior data before applying experimental techniques such as NGS and reanalyzed using new techniques or methods that have not been used in previous studies.

In addition, GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r/) provides a list of genes with statistical differences between the two groups as well as visualization results. The results are presented in form of a table of genes sorted by significance, as well as a collection of graphical plots to aid in the visualization of differentially expressed genes and the assessment of dataset quality. GEO2R uses the Bioconductor project's GEOquery and limma R packages to perform comparisons on the processed data tables provided by the original submitter. Users can define sample groups and assign samples to each group. Visual outputs include the volcano plot, mean difference plot, uniform manifold approximation and projection (UMAP), Venn diagram of the overlap in significant genes between multiple contrasts by groups, boxplot, etc. The table output consists of *p*-values and other statistics used to select reliable genes between the two groups. GEO2R can easily compare the expression levels of the two groups with a few clicks, but it is preferable to use coding and machine learning libraries in R to perform more customized post-analysis.

In this study, the t-test between two groups and the ANOVA technique in more than three groups will be used as statistical analysis techniques. A volcano plot was used to show the difference in expression levels and statistical significance between the two groups, and a heatmap was used to suggest genes that satisfied the threshold. We attempted to identify genes in the OE sample that had expression levels similar to

those in the young sample. Genes that are similar to those found in the young OE sample with were selected, which can be considered to be related to rejuvenation after exercise.

PCA is an unsupervised learning method that takes the total level of gene expression and reduces it to approximately 10 main components. In the form of a scree plot, approximately 10 reduced principal components are presented, and the top four that describe the entire feature well are presented as two-dimensional spaces. The analysis is said to be well performed well if each sample is well grouped by group.

Among ML techniques, a DT is used to provide information and visualization of which factors best describe each group. In this study, a DT was proposed using the R's rpart package of R. Nodes are presented based on the level of expression of a specific gene, and groups based on age and exercise level as each node is followed.

In this section, we will present the results of applying basic statistical analysis, visualization, and decision trees to the GSE8479 and GSE28422 datasets [52,53]. We propose a method for analyzing genes that exhibit similar expression patterns to young samples through exercise during aging. Differences in gene expression levels and differences in gene expression levels according to aging and resistance training were identified by displaying the results of applying a statistical analysis and decision tree through ANOVA for a total of four groups, whether young samples are trained or aged samples are trained. The GEO2R analysis method and results for GSE8479 are also presented, making it possible to obtain prior data from exercise-related datasets more easily.

1) Resistance exercise reverses aging in human skeletal muscle: (GEO dataset: GSE8479)

The purpose of this study was to compare the skeletal muscle samples of healthy older (n=25), younger (n=26), and six-month resistance exercise training program participants of healthy older adults (n=14) [52]. The authors examined whole transcriptomes in each sample before identifying and visualizing transcripts that were expressed differently in older and younger people. Transcriptome data were analyzed using microarray (GPL2700; Sentrix HumanRef-8 Expression BeadChip) and deposited as GSE8479. They provided evidence that exercise reverses a functional decline in older participants, so we looked for reversed patterns of OE participants.

GEO2R allows one to select and visualize genes that differ significantly between the two groups quickly and easily. Three groups of older 25, younger 26, and OE 14 are listed form a total of 65 samples. Younger and OE samples were considered as one group in this review, while older participants were assigned to the remaining groups. We attempted to find a gene that was expressed similarly in the young old samples after exercise. For analysis, press the "Analyze" button (Fig. 8A). GEO2R's volcano plot, mean difference plot, and UMAP are provided (Fig. 8B, C, and D). The expression patterns for each sample can be confirmed using a boxplot that provides the average and distribution ranges for each sample (Fig. 8E). A bar plot was used to visualize one of the genes that was significantly different between the two groups according to the classification method used in this review. It is evident from Fig. 8F that in the USP54 gene, the old sample had more expression patterns than the other two groups (young and OE).

We applied the transcriptome matrix to ML approaches using the two approaches. The first approach was PCA, which reduced 24,353 gene expression information to 10 dimensions. The results of the PCA analysis of the first and second dimensions revealed OE sample-specific clustering (Fig. 9A), whereas young sample-specific clustering was detected in the third and fourth dimensions (Fig. 9B). A scree plot was depicted, and the top three principal components were significant (Fig. 9C). It is believed that further research on genes that change after training is necessary.

The second approach was logistic regression, which is supervised learning. Each subject in the three categories (older, younger, and OE) was divided into 7:3. The model was trained in the training set, which accounted for 70% of the total, and its performance was evaluated in the test set, which accounted for the remaining 30%. The results were presented as a learning curve and confusion matrix, with 12 samples corresponding to the test set in the O, Y, and OE (Fig. 9D and E). Despite the relatively small number of samples in the test set, it was confirmed that all the samples fit the trained model. In the future, we will be able to use this model in more samples and similar studies.

A volcano plot for genes compared to the older group was presented using OE and young as the same group (Fig. 9F). Samples with a *p*-value less than 0.001 and fold change greater than 0.7 or equal to -0.7 were selected. Genes that were relatively highly expressed in the older samples showed a statistically significant pattern overall. Of the 28 filtered genes, 26 genes were more frequently expressed in elderly people, whereas two genes were more frequently expressed in young and OE people. A heatmap gene expression levels with significant differences above a certain level was presented between the two groups (Fig. 9G). The more significant genes are indicated in dark gray in the heatmap's row annotation bar, and the more highly expressed genes are indicated in red. The dis-

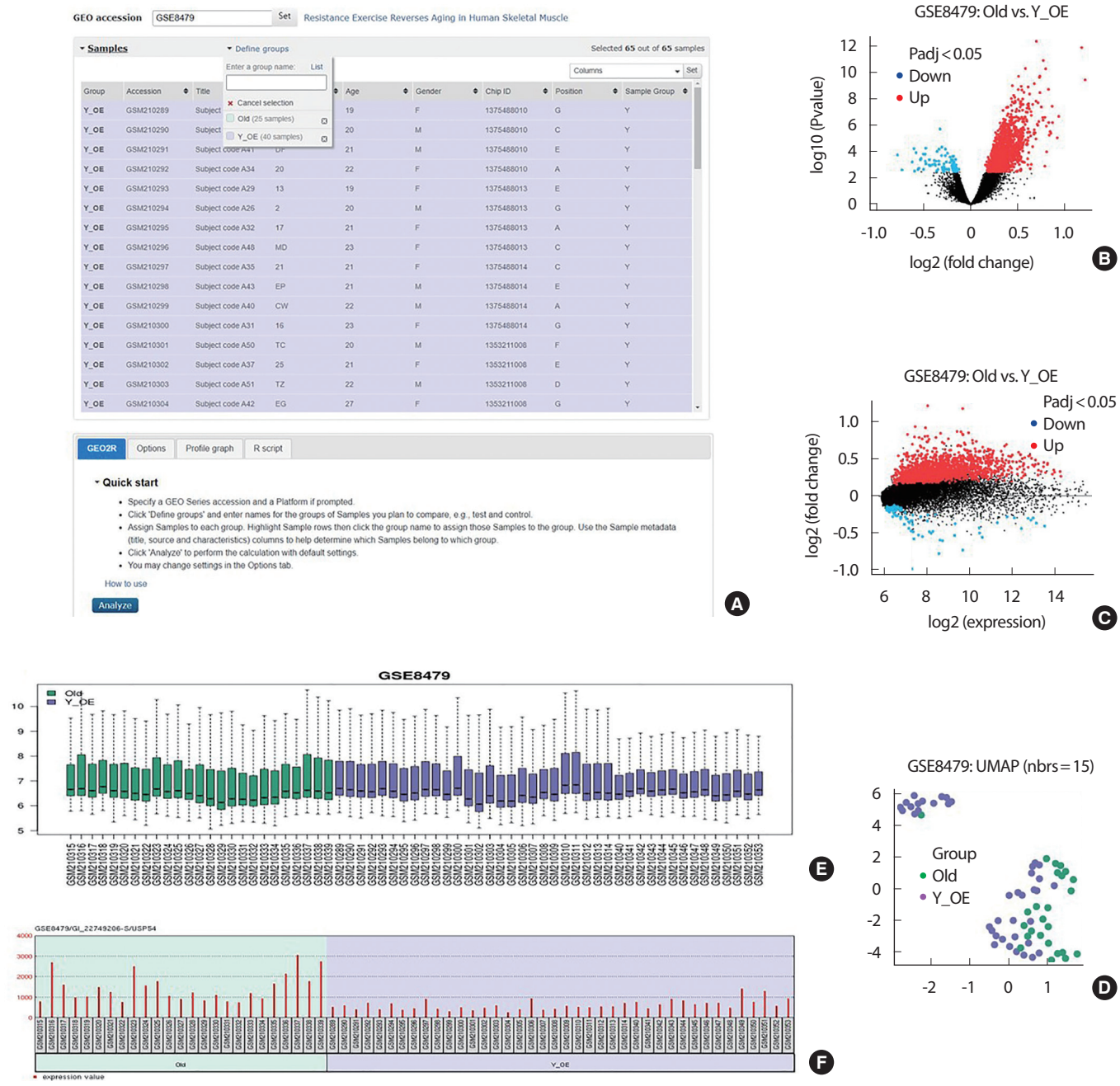


Fig. 8. GEO2R analysis results of the GSE8479 dataset. (A) After selecting each sample in GEO2R, click the set group to assign it. When you click the Analyze button, the analysis proceeds automatically. (B) Volcano plot, (C) mean difference plot, (D) uniform manifold approximation and projection (UMAP) were presented. (E) Boxplot showing the average expression level for each sample, and (F) expression level for each gene can be visualized.

tribution of each sample was presented using the column annotation bar, old samples were classified to the left, and the remaining young and OE samples were classified to the right. Based on the matrix used to create the heatmap, a decision tree was created that explained each of the three groups well, and the three terminal nodes distinguished each class well (Fig. 9H). Based on the expression level of the gene presented in each node, it can be used for research, such as the analysis of physiological changes based on the expression level. It is necessary to confirm cross-validation of genes that are expressed differently in old and young samples, and we expect that they can be used in performance evaluation models related to exercise effects or aging management in the future.

In this way, although it was not discovered in previous studies, recent analysis techniques have been developed and new insights can be drawn by looking at existing data from different aspects.

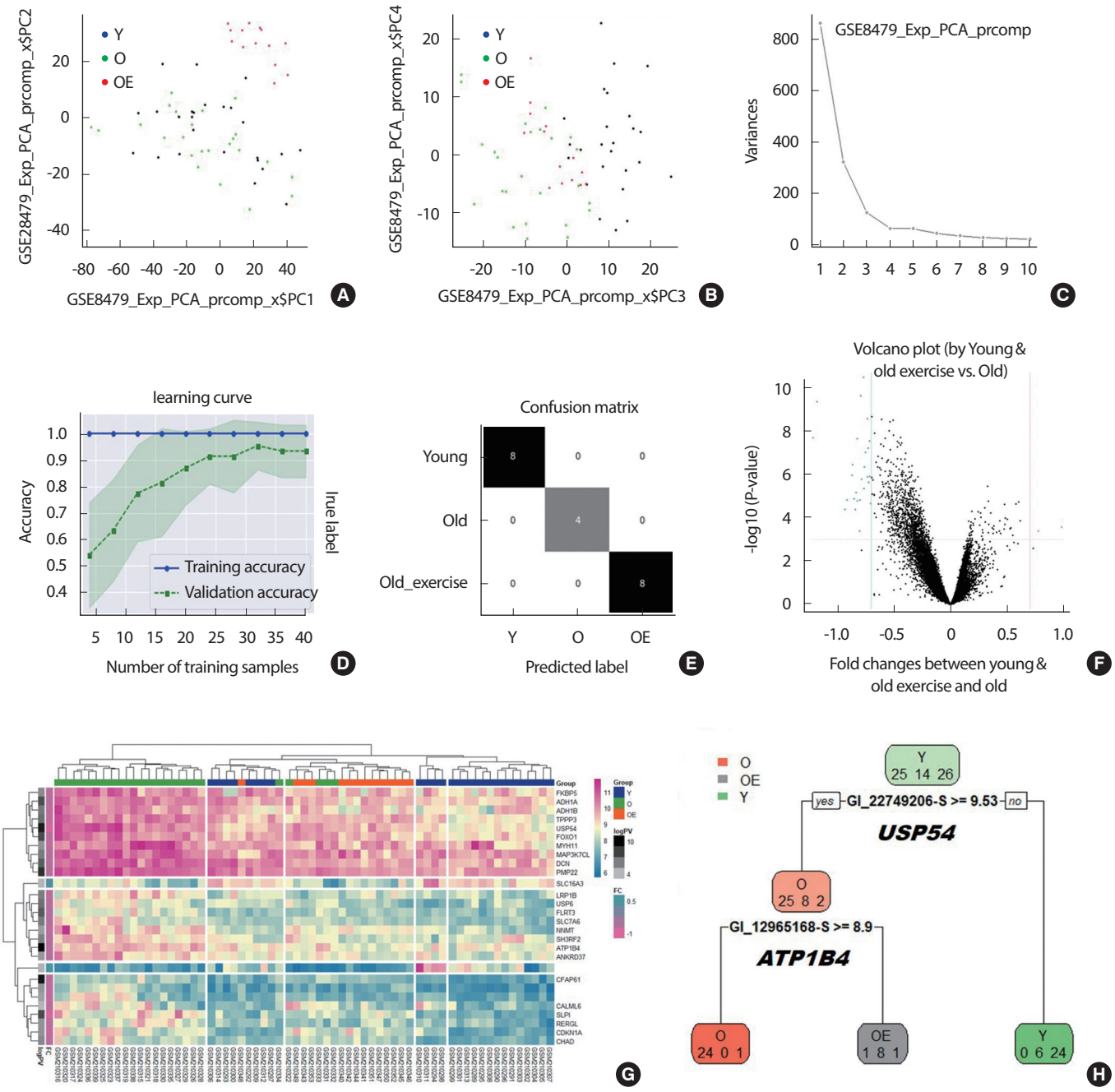


Fig. 9. Analysis results on the GSE8479 dataset. Visualization results for (A) first and second, and (B) third and fourth principal components of expression matrix. (C) A scree plot for the top 10 principal components. (D) The learning curve and (E) the confusion matrix resulting from logistic regression analysis. (F) Exercise-training with older (OE) and young as the same group, volcano plot of genes compared to the older group, and (G) heatmap for the expression levels of genes with significant differences in expression levels above a threshold level between the two groups. (H) Decision tree to classify three groups by gene expression with the three final nodes. Each final node will classify three groups by two gene expression levels.

2) Effects of resistance exercise and resistance training on the skeletal muscle transcriptome in young and old adults: (GEO dataset: GSE28422)

Transcriptome data were analyzed using microarray (GPL570; Affymetrix Human Genome U133 Plus 2.0 Array), and deposited as GSE28422. This dataset was used in the same manner as in a previous study [53], but it was not visualized as a matrix. We selected four categorical variables: age, gender, train state, and time point. Each of the four variables was divided into two categories, resulting in a total of 16 combinations.

The transcriptome matrix was used in two different ML approaches.

https://doi.org/10.15857/ksep.2021.30.2.131

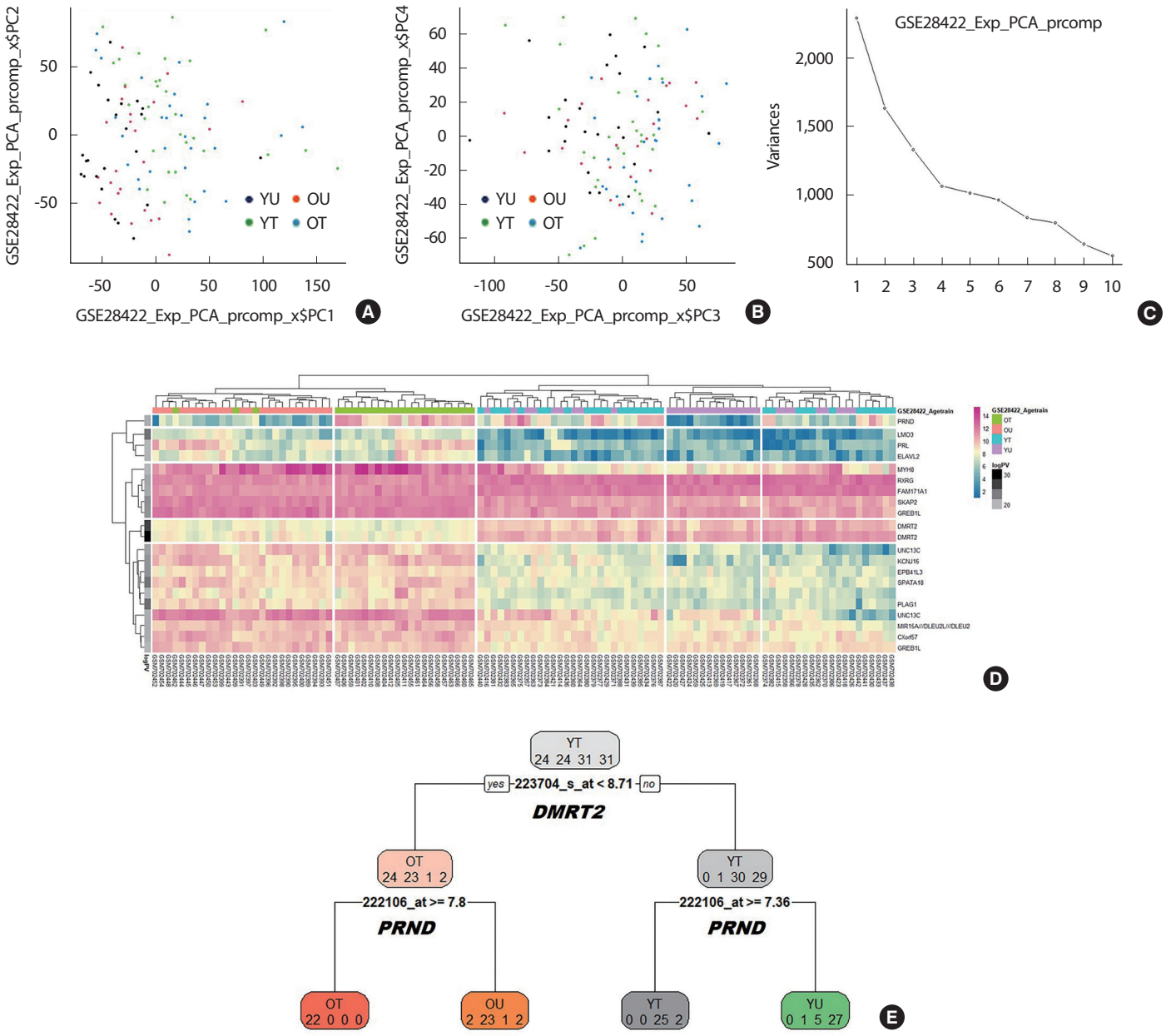


Fig. 10. Analysis results on the GSE28422 dataset. Visualization results for (A) first and second, and (B) third and fourth principal components of expression matrix. (C) A scree plot for the top 10 principal components. (D) Heatmap for the expression levels of genes with significant differences in expression levels between four groups. (E) Decision tree to classify four groups by gene expression with the four final nodes. Each final node will classify three groups by three nodes as two genes.

The first approach was PCA, which reduced a total of 54,674 gene expression information to 10 dimensions. No clustering patterns were found in the results of the PCA analysis of the two dimensions from the first to fourth principal components (Fig. 10A and B). A scree plot was depicted, and the top four principal components were significant (Fig. 10C). Therefore, explaining the four groups using the PCA method proved difficult.

ANOVA was used as the second approach, and genes with different expression levels were selected for each of the four groups. Twenty one genes with a *p*-value less than 10^{-18} were selected. A heatmap of the expression levels of 21 genes was presented between the four groups (Fig. 10D). The more significant genes are indicated in dark gray in the heatmap's row annotation bar. The column annotation bar displayed the sample distribution for each sample. The old sample was divided by

training status, whereas the young sample was not divided. The patterns were divided into four categories: old-trained, old-untrained, young-trained, and young-untrained patterns, as well as genes that were commonly expressed in elderly people. Based on the matrix used to create the heatmap, a decision tree was created that explained each of the four groups well, and the four terminal nodes distinguished each of the four classes well (Fig. 10E). The higher node was classified as old because the expression of the *DMRT2* gene was lower, and the second node was classified as trained in both old and young samples with high *PRND* gene expression. In other words, ML enabled the discovery of two genes that can be classified into two conditions (old versus young and trained versus untrained). In the future, the analysis method described in this review can be used to select genes that adequately explain each condition using ANOVA and decision tree when combining two or more conditions.

3) Sub-conclusion

In this example, we used ML to analyze a single open dataset from the NCBI GEO dataset. This has the advantage of providing new insights through the researcher's individual differentiated approach to data generated in previous studies. Understanding the characteristics of each dataset is critical in this approach.

In the field of exercise science, an attempt was made to create an open dataset so that the change in the transcriptome in human skeletal muscle after exercise could be easily checked [55]. In addition, MetaMEx (http://www.metamex.eu), an online tool that can compare the expression of specific genes in human skeletal muscle based on age, sex, exercise period, exercise type, and biopsy site, has been developed by integrating all human skeletal muscle gene expression datasets uploaded to NCBI GEO [56]. Using this tool, researches can confirm the expression of genes whose relevance to exercise has yet been published [57], and even genes whose functions or roles are known can be used as a preliminary review.

To understand the relationship and correlation between clear biological mechanisms and exercise, sufficient validation in a laboratory environment is required. However, the vast amount of information in the database can be used to develop an *in silico*-based experimental method that can save unnecessary time and money in terms of economics and reduce unnecessary sacrifices of experimental animals in terms of ethics.

Further aspects

In this review, we presented the most recent studies on disease prediction and prognosis using ML technologies. The main objective was to select the appropriate features and classify vital datasets. The future of healthcare solutions can be achieved by sophisticated modeling of drug response data or vital datasets. Advanced modeling of drug response data or vital datasets can pave the way for the future of healthcare solutions. Exercise-related omics datasets will be continuously created. Through an ML-based model that can explain omics datasets, it enables the realization of personalized medicine.

CONCLUSIONS

The prediction models discussed here are ML techniques that use various input characteristics and data samples. Given the growing use of the ML method in biomedical studies, we presented the most recent research using this technology for disease risk or patient outcome modeling. Thus, we expect that these modeling technologies will be applied to exercise-related datasets.

CONFLICT OF INTEREST

The authors have no conflicts of interest relevant to this study.

AUTHOR CONTRIBUTIONS

Conceptualization: K Baek, J Gim; Data curation: K Baek, J Gim; Formal analysis: J Gim; Funding acquisition; Methodology: J Gim ;Project administration; Visualization: J Gim; Writing-original draft: K Baek, J Gim; Writing-review & editing: K Baek, J Gim.

ORCID

Kyung-Wan Baek https://orcid.org/0000-0002-8445-3773
Jung-Jun Park https://orcid.org/0000-0002-2518-7225
Jeong-An Gim https://orcid.org/0000-0001-7292-2520

REFERENCES

1. Hinkson IV, Davidsen TM, Klemm JD, Kerlavage AR, Kibbe WA, et al. A Comprehensive Infrastructure for big data in cancer research: accelerating cancer research and precision medicine. Front Cell Dev Biol. 2017;5:83.

2. Obermeyer Z, Emanuel EJ. Predicting the future-big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216-9.
3. He KY, Ge D, He MM. Big data analytics for genomic medicine. Int J Mol Sci. 2017;18(2)412.
4. Lee CH, Yoon HJ. Medical big data: promise and challenges. Kidney Res Clin Pract. 2017;36(1):3-11.
5. Haghi M, Thurow K, Stoll R. Wearable devices in medical internet of things: scientific research and commercially available devices. Healthc Inform Res. 2017;23(1):4-15.
6. Yang H, Yu J, Zo H, Choi M. User acceptance of wearable devices: an extended perspective of perceived value. Telemat Inform. 2016;33(2): 256-69.
7. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6): 333-51.
8. Sirbu A, Kerr G, Crane M, Ruskin HJ. RNA-Seq vs dual- and single-channel microarray data: sensitivity analysis for differential expression and clustering. PLoS One. 2012;7(12):e50986.
9. Anaissi A, Goyal M, Catchpoole DR, Braytee A, Kennedy PJ. Ensemble feature learning of genomic data using support vector machine. PLoS One. 2016;11(6):e0157330.
10. Bibault JE, Giraud P, Burgun A. Big data and machine learning in radiation oncology: state of the art and future prospects. Cancer Lett. 2016;382(1):110-7.
11. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41(Database issue):D991-5.
12. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, et al. Arrayexpress update--simplifying data submissions. Nucleic Acids Res. 2015;43(Database issue):D1113-6.
13. Hecker M, Ruge A, Putscher E, Boxberger N, Rommer PS, et al. Aberrant expression of alternative splicing variants in multiple sclerosis-a systematic review. Autoimmun Rev. 2019;18(7):721-32.
14. Nakhuda A, Josse AR, Gburcik V, Crossland H, Raymond F, et al. Biomarkers of browning of white adipose tissue and their regulation during exercise- and diet-induced weight loss. Am J Clin Nutr. 2016;104 (3):557-65.
15. Diao JA, Kohane IS, Manrai AK. Biomedical informatics and machine learning for clinical genomics. Hum Mol Genet. 2018;27(R1):R29-34.
16. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321-32.
17. de Bruijne M. Machine learning approaches in medical image analysis: From detection to diagnosis. Med Image Anal. 2016;33:94-7.
18. Grys BT, Lo DS, Sahin N, Kraus OZ, Morris Q, et al. Machine learning and computer vision approaches for phenotypic profiling. J Cell Biol. 2017;216(1):65-71.
19. Mirzaei G, Adeli A, Adeli H. Imaging and machine learning techniques for diagnosis of Alzheimer's disease. Rev Neurosci. 2016;27(8): 857-70.
20. Cho S-B, Won H-H. Machine learning in DNA microarray analysis for cancer classification. in Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics. 2003-Volume 19. 2003.
21. Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Appl Bioinformatics. 2003;2(3 Suppl): S75-83.
22. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, et al. Gene selection from microarray data for cancer classification--a machine learning approach. Comput Biol Chem. 2005;29(1):37-46.
23. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics. 2008;9 Suppl 1:S13.
24. Qin J, Lewis DP, Noble WS. Kernel hierarchical gene clustering from microarray expression data. Bioinformatics. 2003;19(16):2097-104.
25. Horng JT, Wu LC, Liu BJ, Kuo JL, Kuo WH, et al. An expert system to classify microarray gene expression data using gene selection by decision tree. Expert Syst Appl. 2009;36(5):9072-81.
26. Williams-DeVane CR, Reif DM, Hubal EC, Bushel PR, Hudgens EE, et al. Decision tree-based method for integrating gene expression, demographic, and clinical data to determine disease endotypes. BMC Syst Biol. 2013;7(1):119.
27. Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, et al. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomed Signal Process Control. 2019; 52:456-62.
28. Bernard S, Heutte L, Adam S. On the selection of decision trees in random forests. in 2009 International Joint Conference on Neural Networks. 2009. Atlanta, GA, USA: The Institute of Electrical and Electronics Engineers (IEEE).
29. Shinde SR, Bhadikar PS. A genetic algorithm, information gain and artificial neural network based approach for hypertension diagnosis. in 2016 International Conference on Inventive Computation Technologies (ICICT). 2016.

30. Liu X, Wang C, Hu Y, Zeng Z, Bai J, et al. Transfer learning with convolutional neural network for early gastric cancer classification on magnifiying narrow-band imaging images. in 2018 25th IEEE International Conference on Image Processing (ICIP). 2018.
31. Oh SL, Hagiwara Y, Raghavendra U, Yuvaraj R, Arunkumar N, et al. A deep learning approach for parkinson's disease diagnosis from EEG signals. Neural Comput Appl. 2020;32(15):10927-33.
32. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12(10):931-4.
33. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983-7.
34. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535-48 e524.
35. Lee Y, Lee CK. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics. 2003;19(9):1132-9.
36. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1):389-422.
37. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531-7.
38. Moler EJ, Chow ML, Mian IS. Analysis of molecular profile data using generative and discriminative methods. Physiol Genomics. 2000;4(2):109-26.
39. Dai JJ, Lieu L, Rocke D. Dimension reduction for classification with gene expression microarray data. Stat Appl Genet Mol Biol. 2006;5:Article6.
40. Misra J, Schmitt W, Hwang D, Hsiao LL, Gullans S, et al. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. Genome Res. 2002;12(7):1112-20.
41. Musani SK, Zhang HG, Hsu HC, Yi N, Gorman BS, et al. Principal component analysis of quantitative trait loci for immune response to adenovirus in mice. Hereditas. 2006;143(2006):189-97.
42. McVean G. A genealogical interpretation of principal components analysis. PLoS Genet. 2009;5(10):e1000686.
43. Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MG. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 Genomes Data. Mol Biol Evol. 2016;33(4):1082-93.
44. Weintraub WS, Fahed AC, Rumsfeld JS. Translational medicine in the era of big data and machine learning. Circ Res. 2018;123(11):1202-4.
45. Low SK, Zembutsu H, Nakamura Y. Breast cancer: The translation of big genomic data to cancer precision medicine. Cancer Sci. 2018;109(3):497-506.
46. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief Bioinform. 2018;19(2):325-40.
47. Mirza B, Wang W, Wang J, Choi H, Chung NC, et al. Machine learning and integrative analysis of biomedical big data. Genes (Basel). 2019;10(2).
48. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics. 2018;19(4):562-78.
49. Cai J, Luo JW, Wang SL, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018;300:70-9.
50. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. Biophys Rev. 2019;11(1):31-9.
51. Suh C, Fare C, Warren JA, Pyzer-Knapp EO. Evolving the materials genome: how machine learning is gueling the next generation of materials discovery. Annual Review of Materials Research. 2020;50:1-25.
52. Melov S, Tarnopolsky MA, Beckman K, Felkey K, Hubbard A. Resistance exercise reverses aging in human skeletal muscle. PLoS One. 2007;2(5):e465.
53. Raue U, Trappe TA, Estrem ST, Qian HR, Helvering LM, et al. Transcriptome signature of resistance exercise adaptations: mixed muscle and fiber type specific profiles in young and old adults. J Appl Physiol. (1985) 2012;112(10):1625-36.
54. Stewart J, Sprivulis P, Dwivedi G. Artificial intelligence and machine learning in emergency medicine. Emerg Med Australas. 2018;30(6):870-4.
55. Vissing K, Schjerling P. Simplified data access on human skeletal muscle transcriptome responses to differentiated exercise. Sci Data. 2014;1:140041.
56. Pillon NJ, Gabriel BM, Dollet L, Smith JAB, Sardon Puig L, et al. Transcriptomic profiling of skeletal muscle adaptations to exercise and inactivity. Nat Commun. 2020;11(1):470.
57. Baek KW, Yoo JI, Kim JS. Relationship between METTL21C gene expression and exercise in human skeletal muscle: a meta-analyisis. Exerc Sci. 2021;30(1):102-9.