A Deep Learning and Genetic Algorithm Based Feature Selection Processes on Leukemia Data

Maria Frasca (✉ mfrasca@unisa.it)
University of Salerno

Rita Francese
University of Salerno

Michele Risi
University of Salerno

Genoveffa Tortora
University of Salerno

Research Article

Keywords: Microarray, Leukemia, DNN, Pathway, Feature Selection

Posted Date: January 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1247212/v1

RESEARCH

A deep learning and genetic algorithm based feature selection processes on Leukemia Data

Maria Frasca*, Rita Francese, Michele Risi and Genoveffa Tortora

*Correspondence: mfrasca@unisa.it
Dipartimento di Informatica, University of Salerno, Fisciano, Italy
Full list of author information is available at the end of the article

Abstract

Background: One of the challenges in the bioinformatics field is the characterization of genetic diseases, more precisely of the anomalies of the genetic code that lead to the onset of various pathologies. Concerning leukemia, there exist different types, such as acute and chronic leukemia. The acute ones are Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). This paper considers a dataset of patients belonging to two distinct classes: ALL and AML. The aim is to define a feature selection analysis process mainly based on Deep Learning for both classifying the leukemia of patients as ALL or AML and identifying the list of differential expressed genes.

Method: The analyzed data are extracted from dual-channel microarray experiments from the Gene Express Omnibus (GEO) platform, a public database available on the NCBI website containing genomic data, which represent the methylation values for each gene of each sample. The analysis exploits feature selection techniques aimed at reducing the consistent number of variables (genes). To this aim, we use linear models for differential expression for microarray data, and an autoencoder based unsupervised deep learning model to simplify and speed up the classification.

Results: Following the reduction in the number of variables, classification models have been implemented with the use of a deep neural network (DNN), obtaining a classification accuracy of approximately 92%. Then, the results have been compared with the ones provided by an approach based on support vector machines (SVM) giving an accuracy of 87,39%. Moreover, another feature selection approach based on genetic algorithms has been experimented obtaining 60,36% (DNN) and 30,63% (SVM) of accuracy.

Conclusions: For further verification of the relevance of the selected set of genes, we conducted a gene enrichment analysis based on the functional annotation of the differentially expressed genes. As a result, a differentially expressed pathway between the two pathologies has been detected.

Keywords: Microarray; Leukemia; DNN; Pathway; Feature Selection

1 Introduction

Bioinformatics is a multidisciplinary science that analyzes biological, biochemical and biophysical information with computational, mathematical and statistical methods, to formulate hypotheses about life processes. It is one of the scientific areas which get wide benefits from the analysis of big data.

Figure 1 Acute Lymphoblastic Leukemia and Acute Myeloid Leukemia .

Given the heterogeneity of biological information, the different types of experiments and platforms used, as well as the possible presence of noise in the surveys, there is no universally better computational model or technique in terms of performance or accuracy. Thus, it is necessary to define the processes individually, not only for each research area but also for each set of data analyzed. To this aim, different Machine Learning and Deep Learning methodologies have been successfully proposed to detect patterns from bioinformatics big data [1, 2].

Human tumors are characterized by a global loss of DNA methylation associated with hypermethylation of promoters, leading to silencing of the corresponding genes [3]. Usually, hypermethylation and silencing regions contain repetitive elements, which are instead significantly demethylated in cancer cells; while tumor suppressor genes and those involved in DNA repair are silenced with hypermethylation of promoters and can affect tumor response to anticancer drugs [4].

Epidemiological studies aim at correlating epigenetic variations with environmental factors and identifying some diagnostic and prognostic biomarkers that can be used in routine clinical practice [5];

Among the human tumors, leukemia is one of the most relevant: in 2020 worldwide it has been the cause of death for 311,594, while the new leukemia cases have been 474,519[1]. For leukemias we mean a heterogeneous group of neoplastic diseases, which foresee any process of proliferative alteration of a progressive and irreversible nature of the blood cells of the bone marrow. Leukemias originate from the malignant transformation of hematopoietic stem progenitor cells, with alteration of the proliferation and differentiation of the same cells (examples are in Fig. 1).

In leukemias, blasts (i.e., immature and undifferentiated cells) have a proliferative advantage over normal tissue, proliferating uncontrollably. The cells most involved in this process are the white blood cells, also named lymphocytes, which are produced in large quantities by the bone marrow. These leukemia cells could interrupt their maturation process early and then become resistant to programmed death mechanisms, in this way more cells are produced than they die and these, accumulating in the bone marrow, determine an alteration of proliferation and differentiation of normal hematopoietic cells (e.g., red blood cells and platelets).

There exist different types of leukemia. They are commonly divided into acute and chronic, based on the rate of progression of the disease. In acute leukemia, the number of cancer cells increases rapidly and the onset of symptoms is early, while in chronic leukemia the malignant cells tend to proliferate more slowly. Over time, however, chronic forms also become aggressive and cause an increase in leukemia cells in the bloodstream. If the disease arises from the lymphoid cells of the bone marrow (from which white blood cells called lymphocytes develop) it is called Acute Lymphoid Leukemia (ALL), if instead, the starting cell is of the myeloid type (from which red blood cells, platelets and different white blood cells develop from lymphocytes) we speak of Acute Myeloid Leukemia (AML).

[1] https://gco.iarc.fr/today/data/factsheets/cancers/36-Leukaemia-fact-sheet.pdf

In this paper, we define a process aiming at detecting a set of differentially expressed genes in terms of methylation level, i.e., genes that in different conditions have an expression level significantly different in the AML and ALL cases, and their characteristic pathways. The detection of gene expression data samples involves feature selection and classification. To this aim, we adopt Deep Learning models (e.g., feature selection techniques and classifiers methods). The analysis has been performed on a dataset consisting of samples from people with leukemia, characterized by a fixed list of genes; the samples belong to two distinct classes: ALL and AML. The analyzed data are extracted from experiments on the dual-channel microarray (spotted microarrays) Illumina Human Methylation 450k BeadChip. They represent the methylation values for each gene of each sample from the GEO database, developed by the NCBI[2], a heterogeneous resource for data submission and recovery.

The microarray Illumina Human Methylation 450k BeadChip assesses the methylation levels of 485,577 CpG sites, covering 99% of RefSeq genes and the different epigenetically important genomic regions such as CpG island, shore and shelf island, 5' and 3' UTRs, and promoter and gene body [6]. It also quantifies the methylation by treating the DNA with sodium bisulfite; then the DNA converted to bisulfite is subjected to an amplification phase, followed by fragmentation and hybridization with the microarray probes. Following the allele-specific hybridization with a single base extension of the probes, a fluorescent label (ddNTP) is incorporated for detection. The methylation level is determined by the differential signal intensities detected by the two probes.

The process starts with an overview of the data and removal of the batch effect; then pre-processing techniques have been applied to reduce the consistent number of variables (genes) and to simplify and speed up the classification. The reduction in the number of variables (feature selection) is carried out taking into account bioinformatics data extracted from microarray experiments, aimed at classifying genes based on their differential expression between ALL and AML biological states. In particular, for feature selection we apply two different approaches: *i)* the first one, consisting of two steps is carried out by using statistical and artificial intelligence techniques; *ii)* the second one is based on the use of a genetic algorithm. For the classification, two different approaches have experimented: a deep neural network (DNN) and support vector machines (SVM). Finally, we conduct a pathway analysis on the reduced dataset to identify therapeutic targets of leukemia.

The main contributions of the paper are the following:

- the definition and the evaluation of a process for the classification of ALL and AML leukemias based on Bayesian and autoencoder approaches and deep neural networks (DNN);
- the identification of an "RNA degradation" pathway in the reduced features detected by the classification process, as an important factor in the development of ALL leukemia.

The paper is structured as follows: Section 2 introduces the background. Section 3 outlines related works, and Section 4 describes the leukemia dataset. Section 5

[2] https://www.ncbi.nlm.nih.gov/geo

presents the analysis process proposed based on the Bayesian method and autoencoders for feature selection, whilst DNN and SVM for the classification. Section 6 shows the analysis process where feature selection is based on genetic algorithms, while Section 7 discusses the results and their implication in therms of pathway analysis. Finally, Section 8 concludes the paper.

2 Background

Methylation is considered to be biologically important in the pathogenesis of many malignancies, therefore it is extremely important to investigate the methylome (DNA methylation profile) in patients both at diagnosis and as the disease progresses [7].

DNA methylation is a chemical modification of DNA that consists in adding a methyl group mainly in the context of CpG dinucleotides, in which the cytosines of the CpG sites can be methylated to become 5-methylcytosine. Areas of DNA with a high density of CpG sites, small regions about 0.5-2 kb in length, are called CpG islands and are usually located in regulatory regions of constitutive (housekeeping) genes and tissue-expressing genes. specification [8]. The 60% of these segments are located in human gene promoters and less frequently in gene bodies or intergenic regions. Methylation involves the DNA methyltransferase (DNMT) family of enzymes. The DNMT family mainly comprises two types of enzymes: DNMT1, which specifically recognizes methylation sites in a DNA half-strand and copies them to the child strand during replication, ensuring the fidelity of the methylation profile during mitosis; DNMT3a and 3b, which are instead involved in the ”de novo” methylation that occurs during embryonic development and cell differentiation.

Silencing of gene transcription is associated with methylation of CpG sites, and these are located near the transcription initiation sites (TSS) of genes [8]. In cancer cells, these sites can often be methylated, leading to blockage of transcription of many cancer-associated genes. Although methylation of sites contained in the gene body may contribute to a tumor by causing somatic and germline mutations, the function of intergenic CpG methylation is not fully understood. Since this transformation is reversible, the methylated genes can be re-expressed by the use of DNMT inhibitors, such as 5aza-2'-deoxycytidine (Aza-dC) genes [9].

DNA microarray data have great importance in fields such as molecular biology and medicine. Gene expression profiles can provide details to accurately classify cancer samples. This can be used not only for prediction but also for diagnosis, understanding and prognosis of the disease. A microarray dataset consists of a large number of gene expressions. Each expression measures the activity level of genes in a particular tissue, enabling us to compare genes expressed in abnormal cancer tissue with those in normal tissue. DNA microarray analysis is useful for simultaneously studying the expression of thousands of genes, and has been rapidly adopted by the research community for the study of a variety of biological processes. It enables to compare two biological classes to identify the differential expression of genes within them, genes with potential relevance to a wide range of biological processes, including cancer development [10].

We have adopted the standard process of microarray analysis [11] depicted in Fig. 2. For our analysis, we used data from the microarray Human Illumina 450k

Figure 2 Visualization of the process in microarray analysis [11].

Beadchip dataset. It quantifies DNA methylation by treating DNA with sodium bisulfite. The DNA converted to bisulfite is subjected to an amplification step, followed by fragmentation and hybridization to probes on the microarray. The hybridization is allele-specific with a single-base extension of the probes. After this, an out-tag label (ddNTP) is incorporated for detection. The analysis is performed according to the standard protocol provided by Illumina: the DNA is changed to the EZ DNA Methylation kit (Zymo Research), the Bead chip signals are detected and digitized with an Illumina scanner [12]. Bisulfite deaminates unmethylated cytosine, causing its chemical conversion to uracil upon alkaline desulfonation. By selective conversion of cytosine but not 5mC to uracil, followed by PCR and sequencing of cloned amplicon DNA, BGS accurately detects the presence of 5mC in each region of interest at single-nucleotide resolution. After bisulfite conversion, each probe is whole-genome amplified (WGA) and enzymatically fragmented. During hybridization, the WGA-DNA molecules anneal to locus-specific DNA oligomers linked to individual bead types. The two bead types correspond to each CpG locus, one to the methylated (C) and the other to the unmethylated (T) state. Allele-specific primer annealing is followed by single-base extension using DNP-labeled and Biotin-labeled ddNTPs. Both bead types for the same CpG locus will incorporate the same type of labeled nucleotide, determined by the base preceding the interrogated “C” in the CpG locus, and therefore will be detected in the same color channel[3]. The 99% of RefSeq genes are covered, including those in regions of low CpG island density and at risk for being missed by commonly used capture methods[4]. At the end of the process, the chip is scanned to show the intensities of the unmethylated and methylated bead types. The raw data are analyzed and the fluorescence intensity ratios between the two bead types are calculated. A ratio value of 0 represents a non-methylation of the locus; a ratio of 1 concern total methylation; a value of 0.5 means that one copy is methylated and the other is not, in the diploid human genome.

3 Related work

Neural networks are powerful machine learning methods that are often used to learn data representations at multiple levels of abstraction. These representations are useful for many applications such as reconstruction, classification, grouping, and recognition. Prediction models use the neural network capabilities to classify, group samples, or apply statistical analysis [13]. In particular, neural networks are also commonly used to build cancer prediction models from microarray data [13]. The high dimensionality of gene expression profiles is a crucial problem in building these models. To minimize the feature size and maximize the classification performance a feature selection pre-processing phase has to be adopted. It is a type of multi-objective optimization problem.

[3] `https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote_dna_methylation_analysis_infinium.pdf`

[4] `https://cancergenome.nih.gov/abouttcga/aboutdata/platformdesign/illuminamethylation450`

Feature selection on microarray data is an area currently very explored to discriminate a subset of optimal features of the various existing classifiers to obtain maximum accuracy. In the following, we discuss the main results of feature selection approaches applied on microarray data, summarized in Table 1, where we report for each approach the considered datasets and classifiers, and the average accuracy.

Chen *et al.* [14] adopted a Kernel-based clustering methods (KBCGS) for gene selection. They compared the performance of their approach with other algorithms to select an excellent number of features. The Maximum-Minimum Cross-Entropy Criterion [15] is used to determine the best method.

Recently, different approaches have been developed to perform gene selection on a genetic dataset.

Dhrif *et al.* [16] presented a new variant of the Particle Swarm Optimization (PSO) algorithm to increase the classification accuracy and preserve the acceptable dimensions of feature subsets when there are many uninformative data. For this purpose, a new encoding scheme is used for mapping particle positions to probabilities. The aim is to expand the search of features in a continuous space without limiting solutions to local optima. To test the stability and scalability of the algorithm they created synthetic datasets.

Kang *et al.* [17] proposed a relaxed Lasso-Gen (rL-Gen) method for tumor classification in which the dataset is first z-scored normalized, then a relaxed Lasso is applied for gene selection and, finally, a generalized multi-class support vector machine (Gen) is used as a classifier.

Ghosh *et al.* [18] proposed a recursive meta-heuristic model is called Recursive Memetic Algorithm (RMA), inspired by Dawkin's notion of meme. The proposed Recursive Memetic Algorithm (RMA) model improves classification accuracy and has a higher convergence rate in finding the cancer biomarker compared to other meta-heuristics such as genetic algorithm (GA) or basic MA.

Sun *et al.* [19] presented a global feature selection method based on a semidefinite programming model relaxed from the quadratic programming model with maximization of feature relevance and minimization of feature redundancy, i.e., Minimum Redundancy Maximum Relevance (MRMR).

Saini *et al.* [20] proposed a gene masking derived from the genetic algorithm. An optimal gene mask is searched that provides the largest performance gain by removing the largest number of features for the chosen classification algorithm.

Lv *et al.* [21] applied a multi-objective model following the analytic hierarchy process that gives more importance to the detection accuracy than the feature size to build a model such as the multi-objective optimization algorithm (MOEDA). This solution is based on a type of distribution estimation algorithm (EDA) that guides the search for the optimum by building and sampling explicit probabilistic models of promising candidate solutions.

Othman *et al.* [22] proposed and developed multi-objective hybrid cuckoo research with evolutionary operators for gene selection. The evolutionary operators used are double mutation and simple crossover operators. The results of the experiments revealed that the developed algorithm, multi-objective cuckoo search with evolutionary operators, outperformed the cuckoo and multi-objective search algorithms with less significant selected genes.

Table 1 Features selection approaches applied on microarray data.

Dataset	**Approach**	**Classifier**	**Average accuracy**
ALL, AML, DLBCL, Lung, Prostate, Lymphoma, SRBCT, Brain, NCI60	KBCGS [14]	SVM, KNN	93,45%
Leukemia, Prostate, B-cell Lymphoma	PSO [16]	Random Forest	97,22%
DLBCL, CNS, Lung, Ovarian, Brain, Lymphoma, MLL, TOX171	Relaxed Lasso [17]	rL-GenSVM, KNN	96,43%
MLGSE2191, Colon, DLBCL, Leukemia, Prostate, MLL, SRBCT	RMA GA [18]	SVM	95,86%
AML, ALL, Breast, Colon, DLBCL, Lung, Medulloblastoma, Prostate	MRMR [19]	CART, Naive Bayes, Random Forest	83,41%
SRBCT Nearest Centroid Classifier	Gene Masking [20] 100%	Nearest Shrunken Centroid Classifier,	
Leukemia, Colon, DLBCL, Prostate, Wang Breast, Lung Adenocarcinoma, Medulloblastoma	Multi-objective heuristic algorithm [21]	SVM	87,71%
Ovary, Lung, SRBCT, CNS, DLBCL, Prostate, Leukemia	Evolutionary Operators [22]	Wilcoxon	73,60%
Colon, Prostate, Lung	PCA [23]	ANN, GAHI	86,33%
Leukemia, Colon, Lung, Ovarian	Hybrid GA + PSO [24]	ANN	98,63%
Colon, Adenocarcinoma, SRBRT, NCI60	DPCAForest [25]	SVM, Recursive Feature Elimination	90,25%

Cahyaningrum *et al.* [23] proposed a technique based on Principal Components Analysis (PCA) to select the most relevant features. Moreover, they proposed the use of ANN and e GA Hybrid Intelligence (GAHI) for cancer detection. Although ANN is recognized as one of the methods to classify microarray data, GA is used in this case to optimize the ANN architecture.

Wu *et al.* [24] proposed an ANN classifier. To initialize the structure, an algorithm was used to choose input variables on layered links and different activation functions for different nodes. Then, a hybrid method integrating GA and particle swarm optimization (PSO) algorithms were used to identify an optimal structure with the parameters encoded in the classifier.

Deng *et al.* [25] proposed DPCAForest, a deep forest-based model that integrates the deep forest and the component analysis of the dynamic principle. DPCAForest adaptively generates minority samples based on sample distribution and then performs principal component analysis dynamically synchronized with the growth of the deep forest to reveal the important features with the highest variance. With dynamic PCA, the model can perform feature extraction in a data-driven manner based on cross-validation and obtain information on the merging between layers.

Various lines of research use the evolutionary calculus to develop solutions to selection problems. Recently, metaheuristic algorithms have been used to perform genetic selection and their implementation has been studied. However, despite the various methods proposed for genetic selection, they suffer from local and optimal stagnation problems and high computational costs, which therefore cannot guarantee the optimal and reasonable use of metaheuristic algorithms in a wide range of research of identified genes [26].

In our approach, unlike others we use a Bayesian inference method to perform the feature selection on the dataset and autoencoders to perform a second feature selection applied on the genes differentially expressed.

4 Dataset

The examined dataset was extracted by the GEO platform public database containing genomic data, available on the NCBI website[5]. It consists of biomedical data of 556 patients, where 233 were affected by AML and 323 by ALL. Specifically, for each patient, the dataset contains the detected CpG probes described by their methylation value. These data are related to the Infinium Human Methylation 450k BeadChip microarray, a popular technology to explore DNA methylomes [27].

5 The analysis process based on Bayesian and Autoencoders techniques

To perform feature selection we applied two different approaches: (i) Bayesian and autoencoders and (ii) GA. In this section, we describe the former that adopts statistical and artificial intelligence techniques.

Analysis of microarray data is based on the hypothesis that the measured fluorescence intensities are representative of the actual level of expression. The complexity of the microarray experimental protocols makes this technology very variable and sometimes subject to significant systematic distortions. For this reason, some manipulations and transformations are necessary before comparing expression levels to attenuate values affected by random aberrations or systematic variations and

[5] `https://www.ncbi.nlm.nih.gov/gds`

Figure 3 The data analysis process based on Bayesian and Autoencoders feature selection.

maintain all the data on comparable levels. Figure 3 depicts this analysis process consisting of the following six steps.

5.1 Step 1 - Data Cleaning

The considered samples were extracted from the GEO database and were related to the Illumina Human 450k microarray; they are methylation data related to the GPL13534 platform series. Our initial dataset consisted of 556 samples coming from several microarray experiments. The number of probes among the samples is different; to make the samples uniform and to be able to analyze them we standardized the number of probes among the samples by difference, obtaining 451,308 CpG probes per sample. Then, we performed on our dataset a preprocessing phase by removing the cross-reactive probes, the SNPs probes, and the probes related to sex chromosomes.

Cross-reactive probes target repetitive sequences or co-hybridize alternative sequences that are highly homologous to desired targets and therefore spurious signals can be detected. The cross-reactive sites could reflect CpGs of different methylation status or non-CpGs that are detected as fully methylated or unmethylated loci [6]. Equally important is our search for probes that target CpG sites that overlap with $SNPs$ (single nucleotide polymorphism). SNPs are a variation of the genetic material of a single nucleotide, such that the polymorphic allele is present in the population in a proportion greater than 1%. These portions of the genome can interfere with the methylation analyzes and have to be eliminated. We also remove all the probes related to the X and Y chromosomes because we will focus our analysis only on autosomal genes (not related to sex); this is because there is an imbalance of methylation on sex chromosomes. In particular, the X chromosome, inactive in women, is hypermethylated and this would bring noise into the analysis. In this way, the analysis of genes differentially expressed in the two leukemia types is conducted only on autosomal genes. At this point, we obtained a dataset composed of 556 samples and 434,917 CpG probes. The numerical data within the dataset represent the fluorescent intensities of the probes in double-channel microarray experiments. For the i^{th} probe the estimation of the methylation level $\beta_i \in [0, 1]$ is defined as follows [28]:

$$\beta_i = \frac{max(y_{i,methy}, 0)}{max(y_{i,nmethy}, 0) + max(y_{i,methy}, 0) + \alpha}$$

where y_i is the fluorescent intensity of the probe, $methy$, and $unmethy$ are, respectively, the strength of a methylated and unmethylated signal, and α is an arbitrary value (usually 100) used to stabilize β_i values. On these values we performed a gene sets enrichment analysis operation to obtain the related genes: the resulting dataset was composed of 556 samples and 19,340 genes. A further cleanup eliminated the missing values, as they could generate errors in the measurement and understanding of the relationships between the variables, reducing the genes to 17,996.

5.2 Step 2 - Batch Effect Removal

The batch effect is a source of variability that has been added to the samples during manipulation, consisting in the introduction of non-biological variability in an experiment [29]. Many factors contribute to the generation of batch effects, some of these include the type of chip, the platform being analyzed, the laboratory, storage conditions, protocols (sample, amplification, labeling and hybridization), cRNA/cDNA synthesis and conditions of washing. In any case, the batch effects often seriously influence the large-scale automatic processing of genomic data sets.

In this step, the batch effect is identified and removed. First, we identified and evaluated the Genomic Spatial Event (GSE) batch variables, i.e., the type of experiment to which each sample refers, through the correlation between the variables. Batch GSEs were identified by statistical analysis of batch medians. This method compares the distribution of each GSE in a single lot to its distribution in all the other lots using the *Kolmogorov-Smirnov (KS) non-parametric test* that verifies the form of the distributions [30]. The p-values returned by the KS test have been corrected by the False Discovery Rate (FDR). This method considers only the biologically relevant differences in the methylation levels through the absolute difference between the median of all the β_i values within a lot for a specific GSE and the respective median of the same GSE in all the other lots. GSEs that had a p-value of significance corrected for FDR lower than 0.01 and had a median difference greater than 0.05 were considered as GSE "batch". After identifying the individual GSE batches, we evaluated the importance of the batch effect in individual batches by considering the number of batch GSEs in the batch and the extent of the deviation of the batch GSE medians in a lot compared to all other lots. We deleted the batch effect on the GSE by using an empirical Bayesian framework and then we validated the results: no further effect was detected.

5.3 Step 3 - Normalization

The sample data have been normalized to remove systematic variation in a microarray experiment that affects the measured gene expression levels. One of the objectives of DNA microarray analysis is to compare the levels of gene expression in two or more pathological conditions to identify their peculiarities. For our dataset, we have adopted quantile normalization, whose aim is to make equal the empirical intensity distributions of all arrays.

The quantile normalization transforms the intensity distributions of each specific array. In particular, it assigns to each intensity the same value to the quantile to which it belongs. Thus, each intensity has the same distribution in all arrays. This method is based on the consideration that a quantile-quantile graph is a line perfectly coinciding with the diagonal if and only if the distribution of the two data vectors is the same [31].

This means that it is possible to give all arrays the same distribution by replacing the values of the original dataset with the average quantile by applying the following normalization algorithm.

Let M be a matrix of ng genes (rows) and n arrays (columns) representing the number of patients:

1. Sort each column of M, obtaining M_{sort};

2. to each element of the k_{th} row in X_{sort} assign the average value of that row, obtaining M'_{sort};
3. calculate $M_{normalized}$ by reordering each column of M'_{sort} according to the original order.

A negative aspect of this method is that it forces the quantile values to be all the same. This could be a problem in the distribution queues, where it is possible for a gene to have the same value on all arrays, even if this situation rarely occurs [31].

5.4 Step 4 - Bayesian Feature Selection

In this step, we identified the differentially expressed genes between the two pathological classes AML and ALL by using a Bayesian feature selection technique. Bayesian methods are suitable to study multidimensional inference problems, so it naturally applies to microarray data [32]. Unlike the methods that apply classical inference separately for each gene, the Bayesian analysis exploits information sharing between the genes. We adopted Limma (Linear Models for MicroArray data)[6] to identify these differentially expressed genes through the use of the empirical Bayesian method.

In the following, we describe the adopted feature selection procedure.

Two matrices are obtained from the dataset: the design matrix, containing the samples in the array, and the contrast matrix, which specifies the comparisons to be performed on the samples. The design matrix is specified as follows:

1. the rows represent samples;
2. the columns represent groups. In our case, there exist two columns representing ALL and AML samples, respectively;
3. for each sample, the column corresponding to its group has a coefficient equal to 1, otherwise 0.

The contrast matrix represents the difference between the columns.

A first fitting function is applied to the design matrix and data from an experiment involving a series of microarrays with the same set of probes. The function adopts multiple linear models for weighted or generalized minimum squares. Thus, the linear model is adapted to the expression data (by gene) for each probe. A second fitting function is applied to the output of the first regression and the contrast matrix. Given the linear model previously computed, the fitting function calculates the estimated coefficients and standard errors for a given set of contrasts. The function re-orientates the adapted model from the original matrix design to any set of contrasts of the original coefficients. We then applied an empirical Bayesian model for linear data regression, which dynamically borrows information between genes. This function is used to classify genes in order of evidence based on their differential expression; the fact that the same linear model is adapted to each gene allows us to borrow the relationships between genes to moderate residual variances.

Bayesian inference is an approach to statistical inference in which probabilities are not interpreted as frequencies, proportions or similar concepts, but rather as levels of confidence in the occurrence of a given event. The Bayesian model has been very successfully applied in gene expression analyses to moderate the variance

[6] `https://bioconductor.org/packages/release/bioc/html/limma.html`

estimators gene-wise, and furthermore, in Limma the estimate of global variance can incorporate a tendency to average variance.

We get an estimate of the prior distribution from the marginal distribution of the observed data [33].

The degrees of freedom for the individual variances have increased to reflect the extra information obtained from Bayes' empirical moderation, resulting in an increase in statistical power to detect differential expression [34].

A table of top-ranked genes from the adapted linear model is extracted. This table contains various summary statistics for the top-ranked genes and the selected contrast. In particular, it contains the variables *logFC* and *adj.P.Val*, adopted to perform the feature selection. *logFC* provides the value of the contrast; usually, this represents a change of log_2 times between two or more experimental conditions, while *adj.P.Val* is the distribution of *p-values* adjusted by Benjamini-Hochberg correction [35], which introduced FDR, i.e., the expected proportion of the number of false-positive results on the total of all the positive results, and represents the number of null hypotheses wrongly refused on the total of those rejected. For this reason, we initially extracted the genes with a value of $adj.P.Val < 0.01$ and with a $|logFC| > 2$, thus obtaining 1,118 genes for 556 samples.

5.5 Step 5 - Feature Selection with Autoencoders

To further reduce our features we used autoencoders. An autoencoder is an unsupervised artificial feed-forward neural network. Conceptually, it is similar to PCA and can be used to reduce the dimensional space.

The autoencoders compress the input data by forcing the network to use a low-dimensional representation of data. They can to reconstruct the original input. An autoencoder consists of 3 layers: an input layer, a hidden layer and an output layer. The number n of input nodes of this type of network is equal to the number of output nodes (the number of genes). In our case, the hidden layer is composed of two nodes as shown in Fig. 5.5. An autoencoder is divided into two parts: the encoder that learns the mapping between the unlabeled high-dimensional $I[1 : n]$ input data and the low-dimensional representations (in the bottleneck layer), and the decoder that learns the mapping from the intermediate layer representation to the output reconstructed in a high dimension $O[1 : n]$. Autoencoder-based approaches learn to reconstruct input samples by optimizing the Root Mean Squared Error (RMSE) objective function [36].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(I\left[i\right] - O\left[i\right]\right)^2}{n}} \tag{1}$$

For autoencoders, we have chosen the batch size equal to the number of genes in our dataset (1,118 genes), Adam's optimization type, an algorithm for the optimization of the gradient of the first order of the stochastic objective functions, based on adaptive estimates of moments of lower order [37]. As activation functions we used "$tanh$", "$linear$", "$tanh$", respectively for the three layers. At the end of the process we obtained 28 abnormal genes that we removed, obtaining a final dataset composed of 1,090 genes.

Figure 4 The DNN architecture, where n represents the number of genes.

5.6 Step 6 - Classification

For the classification, we experimented two different techniques: a deep neural network (DNN) and support vector machines (SVM), as described in the following.

5.6.1 DNN

We adopted the DNN shown in Fig. 4. The size of the input is equal to the number of genes. The input layer is composed of 30 units with the "*relu*" activation function. The 4 hidden layers transform the representation of the previous layer into a more abstract form. They are composed of 22, 15, 9, 5 respectively, with "*relu*" activation function. The final classification is performed by the last layer (the output layer), composed of 2 units with the "*softmax*" activation function. To classify the class of patients (i.e., AML or ALL) we adopted the "*categorical cross entropy*" loss function. To set up the hyperparameters we selected the "*adam*" optimizer. We adopted the k-fold cross-validation, with $k = 5$.

Figure 5 shows the average loss and average accuracy results on the training and validation sets (blue line and green line, respectively). For the test set we achieved 0,2499 of loss and 91.89% of accuracy.

5.6.2 SVM

We also experimented the use of a support vector machine (SVM) [38] as a classifier. SVM are supervised learning models and a powerful technique for classification and regression, with associated learning algorithms. Given a set of training examples, each one belonging to one of two different categories, a training algorithm creates a model that assigns new examples to one or the other category; the model is then used to make predictions for a set of test examples. We got an accuracy of 87,39% by using the same training and test sets adopted with the DNN.

6 The analysis process based on a Genetic Algorithm

In this section, we present the second feature selection process we experimented with, based on a genetic algorithm (GA), widely used for this purpose in the literature (e.g., [39, 40]). In particular, we followed the analysis process shown in Fig. 6 that differs from the one in Fig. 3 for the red step. Therefore, in the following, we describe only the Feature Selection Genetic Algorithm step and the final classification results.

GAs are heuristic adaptive search algorithms for solving research and optimization problems. They follow a heuristic process (depicted in Fig. 7) inspired by the

Figure 5 Loss and accuracy results of the applied neural network on the feature selection implemented with Limma and autoencoders, where "val"/"$loss$" represents the average accuracy/loss of the training set, and "val_acc"/"val_loss" represents the average accuracy/loss of the validation set, respectively.

Figure 6 The data analysis process based on Genetic Algorithm feature selection.

genetics and the principle of natural selection by Charles Darwin. GAs use a set of solutions that evolves at intervals called generations. The evolution is guided by the search for the optimal solution using comparison. In particular, a fitness function is adopted for selecting the best individuals of the current generation that will be used to create the next generation [41]. GA involves a cyclical operation that simulates the evolutionary process of a population. Each cycle represents a generation and consists of operations carried out to generate a new population made up of increasingly better individuals.

Our dataset consists of 16,408 individuals (i.e., genes). We adopted a GA for feature selection starting from a binary array that represents a chromosome, where genes are the array elements. A chromosome is generally encoded with a bit or character vector. In our case, each element is set to 1 if the biological gene is not expressed (i.e., its value is less equal than 0,5), 0 otherwise. The population is a set of solutions (chromosomes) related to the considered problem.

In the following, we describe the adopted GA, instantiated with the parameters in Table 2.

The first step of this algorithm creates the initial population (i.e., 100 individuals) randomly setting the binary values of genes, while the next phases are repeated with each generation and are associated with the principle of natural selection or genetics.

In GAs, individuals have also named chromosomes because of their structure and operations defined on them. Each solution is described by a set of characteristics very similar to the genes and new solutions are created by applying the same mutation and crossover operators present in genetics [42]. The selection of the best individuals is performed by combining or modifying the characteristics that identify an individual. From genetics, the new chromosomes are obtained by recombining their genetic heritage or by changing the genes with the mutation and crossover operators. For each combination of genes, it is possible to calculate a value called fitness which indicates the ability with which the chromosome or solution can solve the problem. In natural selection, this value measures the individual's adaptation to the environment. So, a better fitness is linked to a greater probability of survival,

Figure 7 Steps of the GA for feature selection.

Table 2 GA parameters.

Parameter	Value
Population size	100
Number of generations	100
Fold for cross validation	5
Crossover probability	0.8
Probability of mutation	0.1
Independent probability of crossover	0.8
Independent probability of mutation	0.08
Tournament size	3

Table 3 Comparison with the adopted features selection approaches.

Dataset size	Approach	Obtained dataset size	Classifier	Accuracy
17.996×556	Limma + autoencoders	1.090×556	DNN	91.86%
			SVM	87,39%
17.996×556	Genetic Agorithm	770×556	DNN	60,36%
			SVM	30,63%

while within the genetic algorithm to a greater probability of selection. Genetic algorithms are stochastic algorithms in which randomness plays an essential role: both phases of selection and reproduction need procedures involving randomness [43]. Concerning the third step (selection), which favors the selection of better individuals of a population to influence the next generation, we adopted a tournament mechanism. A small number of individuals (i.e., *tournament size = 3*) is choose randomly with replacement. We keep the fittest one. This is done again and again until you have got 100 individuals.

The fitness function resolves an optimization problem by maximizing the cross-validation accuracy score with the minimum number of selected genes. The score is the accuracy of training data using only the values of the selected genes. In particular, we divided every dataset into 5 equal parts to calculate the fitness value. Then, we selected one of the mentioned parts as a test set and the rest as a training set. We repeated this action five times for every separate part.

In the Crossover phase, the generation of offspring occurs starting from the parents previously selected in the selection phase. The Crossover operator randomly selects a pair of individuals from the pool of solutions for reproduction, with crossover probability (i.e., 0.8); the values of the two solutions are exchanged to generate two new solutions (i.e., the offspring). Crossover aims to generate two new solutions starting from the combination of two previous ones. The crossover probability determines if crossover will happen. A randomly generated floating-point value is compared to this probability, and crossover is performed if the value is less than that probability; otherwise, the offspring is identical to the parents. Moreover, the independent Crossover probability (set to 0,8) concerns the possibility to select a specific gene to perform the exchange between two parents.

The Mutation phase creates a new population starting from the solutions identified in the previous step. Mutation aims to prevent the locking up at a local minimum and to explore the entire research space when each individual in the population reaches a level of fitness close to the average. It mainly maintains genetic diversity within the population. This operator randomly flips (i.e., one becomes zero or vice-versa) some elements of the offspring. Like crossover, there is a mutation probability (set to 0,1). If a randomly selected floating-point value is less than the mutation probability, the mutation is performed on the offspring; otherwise, no mutation occurs. The mutation is performed by randomly selecting (with an independent mutation probability set to 0,08) a gene in the offspring's chromosome and generating a new value uncorrelated to the previous one.

The GA algorithms repeat the process from step 2 until the number of generations (i.e., 100) is reached.

As output, we tried to obtain a better set of individuals that, with the advancement of generations, contains the subsets of genes involved in both AML and ALL.

Figure 8 The "RNA degradation" pathway [44].

In particular, we selected the genes set to 1. We obtained a reduced dataset containing 6,240 features (genes).

It is worth mentioning the limitations of genetic algorithms. Like most stochastic methods, they do not guarantee success in finding the overall optimal solution to a problem but are often "acceptable" solutions. GAs differ from traditional optimization techniques for several reasons. One of them is that traditional algorithms perform the search starting from a single point while genetic algorithms operate on an entire population of points, and therefore of solutions. This helps in terms of algorithm robustness as it increases the chances of reaching the global optimum and reduces the chances of getting stuck at certain points. For the classification, we used the DNN and the SVM as we did previously for the first process. By applying the DNN to the GA results we obtained 60,36% in accuracy and loss of 0,671. While applying the SVM on the same data we obtained an accuracy of 19,96%.

7 Results and implications

In this section, we compare the results of the processes described in Section 5, we examine the relevance of the selected genes by applying the pathway analysis, and further validate the results.

7.1 Comparison

The obtained classification results are summarized in Table 3, where we reported for each adopted feature selection process (Limma and the autoencoders vs. GA), the number of selected genes and the classification results obtained with both DNN and SVM classifiers. Results show that feature selection using Limma and the autoencoders performs better with both the classifiers, but the deep neural network (DNN) reached higher accuracy (91,86%) than SVM (87,39%).

7.2 Applying pathways analysis

Pathways analysis provides a mean to map key biological processes into important clinical features in disease [45]. It is mainly adopted for predicting cancer outcomes through genome-wide characterizations. In this section, we describe the pathway analysis conducted on the results of the feature selection process based on Limma and autoencoders that reached the best accuracy results and reduced the gene numbers from 19,340 to 1,090. For confirming the relevance at biological level of these results we performed the Gene Sets Enrichment Analysis (GSEA) for the interpretation of gene expression data, which highlights groups of genes that share biological functions (i.e., pathway), chromosomal position or common regulation [46]. In this analysis, we referred to the Kyoto Encyclopedia of Genomes (KEGG)[7], one of the most used databases for pathway knowledge. It links genomic information with functional information of a higher order, computerizing current knowledge on cellular processes and standardizing the genetic annotations [47]. The procedure[8] takes

[7] `https://www.genome.jp/kegg`

[8] `https://bioconductor.org/packages/release/bioc/html/missMethyl.html`

Figure 9 Results of PCA: components ?? and scores ??.

Figure 10 Gene enriched from "RNA degradation" pathway.

a character vector of significant CpG sites, maps the CpG sites to Entrez Gene IDs to test for GO or KEGG pathway enriched using a hypergeometric test, taking into account the number of CpG sites for gene on EPIC array. In particular, statistical approaches to identify significantly overexpressed CpG groups, by examining p-value and FDR are used. Finally, we have extracted only the pathways with $FDR < 0.05$ [48]. As a result, the analysis detected the "RNA degradation" pathway (see Fig. 8) from 20 genes of the 1,090 differentially methylated genes individuated by the feature selection process based on autoencoders. Fig. 10 lists the pathway's detected genes.

RNA degradation in eukaryotic cells plays a very important role in gene expression, as it balances the transcription rate and also serves to rapidly eliminate transcriptions that are no longer needed. Furthermore, RNA degradation plays a controlling role by eliminating RNA molecules that are considered non-functional or abnormal if they lack sequences or characteristic changes necessary for their functions.

The detected pathway denotes the deregulation of transcription, which is an important factor in the development of leukemia. In particular, in T-cell acute lymphoblastic leukemia (T-ALL) it identifies mutations in the RNA decay factors, including mutations in the CNOT3 gene, which is part of the CCR4-NOT complex that regulates gene expression transcriptionally and post-transcriptionally [49]. This gene is included in the 1,090 we detected and that seems to be involved in mRNA deadenylation. When errors occur in this process, there are quality control mechanisms that detect and eliminate defective transcripts that can lead to dysfunctional or toxic protein. However, these mechanisms do not only ensure the fidelity of RNA transcripts but also perform important regulatory tasks by allowing rapid modulation of steady-state RNA levels in response to changes in the intracellular or extracellular environment [50]. However, it remains unclear how mutations in RNA processing may contribute to the development of leukemia [51].

7.3 Further Validation

To further validate the process with best results, e.g. that in Fig. 3, we experimented with the use of Principal Component Analysis (PCA) [52, 53] in substitution of the *Feature Selection with Autoencoders* step. In particular, PCA aims at replacing p (more or less correlated) variables with $k \leq p$ uncorrelated linear combinations (projections) of the original variables. These k principal components are ranked in order of importance by their explained variance, and each variable contributes to each component to varying degrees. The criterion of greatest variance may be similar to feature extraction, where the principal components are used as new features instead of the original variables [52].

Main components and related scores are shown in Fig. 9**??** and Fig. 9**??**, respectively. As it is easy to note, Component 1 was the most characterizing (see Fig. 9**??**). Thus, we isolated this component. Considering the score values of Component 1, we built the graph depicted in Fig. 9**??** obtaining the significance degree for each gene. Then, we removed from the dataset all the genes with a score value in the range $[-6, +6]$, as suggested by the graph, thus reducing the dataset to 1,094 genes for 556 samples. The selected genes in Component 1 had a score greater equal than 0,50 (e.g., genes with strong or moderate loading factor) [54]. Let us note that the genes selected by using PCA are not the same as the ones obtained by the approach based on Limma and autoencoders: the genes in common are 1,068. Thus, the two approaches differ only in 26 genes.

The same procedure previously adopted for the pathway analysis was also applied to the dataset obtained with PCA. We analyzed the list of the genes in the "RNA degradation" pathway and the genes selected by the PCA. We observed that PCA retrieves only 3 genes belonging to the considered pathway. Moreover, PCA requires that information is standardized before its application. Despite the adopted dataset was standardized as described in Section 5 by cleaning the data and by eliminating the batch effect the pathway was not detected. This may be because PCA considers only linear relationships and it does not take into account the potential multivariate nature of biological data.

8 Conclusion

In this paper, we presented a process to detect a set of differentially expressed genes and a pathway in leukemia by adopting feature selection techniques and classifier methods. The analysis has been performed on a dataset consisting of samples from people belonging to ALL and AML classes of leukemia. The classification models have been implemented by using a neural network obtaining a classification accuracy of approximately 92%. We have also detected a set of genes that seems to be involved in leukemia onset. This result is important because pathways analysis provides a means to map key biological processes into important clinical features in disease. We also experimented with the use of a genetic algorithm but with worst results. Another method largely adopted for feature selection, PCA, was also assessed. Results revealed that PCA failed in detecting genes useful for predicting leukemia onset.

Declarations

Ethics approval and consent to participate

This study aimed to mine data for bioinformatics analysis and did not involve original animal and human clinical experiments, so ethical review was not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The dataset analyzed in this paper is a public dataset available at: `https://www.ncbi.nlm.nih.gov/geo`.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

The authors Maria Frasca, Rita Francese, Michele Risi, and Genoveffa Tortora contributed to the manuscript equally.

Author details

Dipartimento di Informatica, University of Salerno, Fisciano, Italy.

References

1. Kashyap, H., Ahmed, H.A., Hoque, N., Roy, S., Bhattacharyya, D.K.: Big data analytics in bioinformatics: architectures, techniques, tools and issues. Network Modeling Analysis in Health Informatics and Bioinformatics **5**(1), 1–28 (2016)
2. Min, S., Lee, B., Yoon, S.: Deep learning in bioinformatics. Briefings in Bioinformatics **18**(5), 851–869 (2017)
3. Miousse, I.R., Koturbash, I.: The fine LINE: methylation drawing the cancer landscape. BioMed Research International **2015** (2015)
4. Sandoval, J., Esteller, M.: Cancer epigenomics: beyond genomics. Current Opinion in Genetics & Development **22**(1), 50–55 (2012)
5. Barrow, T.M., Michels, K.B.: Epigenetic epidemiology of cancer. Biochemical and Biophysical Research Communications **455**(1-2), 70–83 (2014)
6. Chen, Y.-a., Lemire, M., Choufani, S., Butcher, D.T., Grafodatskaya, D., Zanke, B.W., Gallinger, S., Hudson, T.J., Weksberg, R.: Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics **8**(2), 203–209 (2013)
7. Heller, G., Topakian, T., Altenberger, C., Cerny-Reiterer, S., Herndlhofer, S., Ziegler, B., Datlinger, P., Byrgazov, K., Bock, C., Mannhalter, C., *et al.*: Next-generation sequencing identifies major DNA methylation changes during progression of Ph+ chronic myeloid leukemia. Leukemia **30**(9), 1861–1868 (2016)
8. Jones, P.A.: Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics **13**(7), 484–492 (2012)
9. Navada, S.C., Steinmann, J., Lübbert, M., Silverman, L.R., *et al.*: Clinical development of demethylating agents in hematology. The Journal of Clinical Investigation **124**(1), 40–46 (2014)
10. Quackenbush, J.: Microarray analysis and tumor classification. New England Journal of Medicine **354**(23), 2463–2472 (2006)
11. White, C.A., Salamonsen, L.A.: A guide to issues in microarray analysis: application to endometrial biology. Reproduction **130**(1), 1–13 (2005)
12. Darst, R.P., Pardo, C.E., Ai, L., Brown, K.D., Kladde, M.P.: Bisulfite sequencing of DNA. Current Protocols in Molecular Biology, 7–9 (2010)
13. Daoud, M., Mayo, M.: A survey of neural network-based cancer prediction models from microarray data. Artificial Intelligence in Medicine **97**, 204–214 (2019)
14. Chen, H., Zhang, Y., Gutman, I.: A kernel-based clustering method for gene selection with gene expression data. Journal of Biomedical Informatics **62**, 12–20 (2016)
15. Mohammadi, M., Noghabi, H.S., Hodtani, G.A., Mashhadi, H.R.: Robust and stable gene selection via maximum–minimum correntropy criterion. Genomics **107**(2-3), 83–87 (2016)
16. Dhrif, H., Giraldo, L.G.S., Kubat, M., Wuchty, S.: A stable combinatorial particle swarm optimization for scalable feature selection in gene expression data. arXiv preprint arXiv:1901.08619 (2019)
17. Kang, C., Huo, Y., Xin, L., Tian, B., Yu, B.: Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. Journal of Theoretical Biology **463**, 77–91 (2019)
18. Ghosh, M., Begum, S., Sarkar, R., Chakraborty, D., Maulik, U.: Recursive memetic algorithm for gene selection in microarray data. Expert Systems with Applications **116**, 172–185 (2019)
19. Sun, S., Peng, Q., Zhang, X.: Global feature selection from microarray data using lagrange multipliers. Knowledge-Based Systems **110**, 267–274 (2016)
20. Saini, H., Lal, S.P., Naidu, V.V., Pickering, V.W., Singh, G., Tsunoda, T., Sharma, A.: Gene masking-a technique to improve accuracy for cancer classification with high dimensionality in microarray data. BMC Medical Genomics **9**(3), 261–269 (2016)
21. Lv, J., Peng, Q., Chen, X., Sun, Z.: A multi-objective heuristic algorithm for gene expression microarray data classification. Expert Systems with Applications **59**, 13–19 (2016)
22. Othman, M.S., Kumaran, S.R., Yusuf, L.M.: Gene selection using hybrid multi-objective Cuckoo search algorithm with evolutionary operators for cancer microarray data. IEEE Access **8**, 186348–186361 (2020)
23. Cahyaningrum, K., Astuti, W., *et al.*: Microarray gene expression classification for cancer detection using artificial neural networks and genetic algorithm hybrid intelligence. In: Procs. of the International Conference on Data Science and Its Applications (ICoDSA), pp. 1–7 (2020). IEEE
24. Wu, P., Wang, D.: Classification of a DNA microarray for diagnosing cancer using a complex network based method. IEEE/ACM Transactions on Computational Biology and Bioinformatics **16**(3), 801–808 (2018)
25. Deng, X., Xu, Y.: Cancer classification using microarray data by DPCAForest. In: Procs. of the International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1081–1087 (2019). IEEE
26. Jahwar, A., Ahmed, N.: Swarm intelligence algorithms in gene selection profile based on classification of microarray data: a review. Journal of Applied Science and Technology Trends **2**(01), 01–09 (2021)
27. Dedeurwaerder, S., Defrance, M., Bizet, M., Calonne, E., Bontempi, G., Fuks, F.: A comprehensive overview of Infinium HumanMethylation450 data processing. Briefings in Bioinformatics **15**(6), 929–941 (2014)
28. Xie, C., Leung, Y.-K., Chen, A., Long, D.-X., Hoyo, C., Ho, S.-M.: Differential methylation values in differential methylation analysis. Bioinformatics **35**(7), 1094–1097 (2019)
29. Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics **8**(1), 118–127 (2007)
30. Akulenko, R., Merl, M., Helms, V.: BEclear: batch effect detection and adjustment in DNA methylation data. PloS One **11**(8), 1–17 (2016)

31. Bolstad, B.M., Irizarry, R.A., Åstrand, M., Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19**(2), 185–193 (2003)
32. Fisher, C.K., Mehta, P.: Bayesian feature selection for high-dimensional linear regression via the Ising approximation with applications to genomics. Bioinformatics **31**(11), 1754–1761 (2015)
33. Phipson, B., Lee, S., Majewski, I.J., Alexander, W.S., Smyth, G.K.: Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. The Annals of Applied Statistics **10**(2), 946 (2016)
34. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K.: Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research **43**(7), 47–47 (2015)
35. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: series B (Methodological) **57**(1), 289–300 (1995)
36. Pumsirirat, A., Yan, L.: Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. International Journal of Advanced Computer Science and Applications **9**(1), 18–25 (2018)
37. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
38. Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al.: A practical guide to support vector classification. Taipei (2003)
39. Baur, B., Bozdag, S.: A feature selection algorithm to compute gene centric methylation from probe level methylation data. PloS One **11**(2), 1–19 (2016)
40. Sayed, S., Nassef, M., Badr, A., Farag, I.: A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. Expert Systems with Applications **121**, 233–243 (2019)
41. Schaffer, J.D., Grefenstette, J.J.: Multi-objective learning via genetic algorithms. In: Procs. of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 593–595. Morgan Kaufmann Publishers Inc., ??? (1985)
42. Sivanandam, S.N., Deepa, S.N.: Introduction to Genetic Algorithms. Springer, ??? (2008)
43. Mirjalili, S.: Genetic algorithm. In: Evolutionary Algorithms and Neural Networks, pp. 43–55. Springer, ??? (2019)
44. The RNA degradation pathway. `https://www.genome.jp/kegg-bin/show\_pathway?ko03018`. [Online]
45. Efroni, S., Schaefer, C.F., Buetow, K.H.: Identification of key processes underlying cancer phenotypes using biologic pathway analysis. PloS One **2**(5), 425 (2007)
46. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Procs. of the National Academy of Sciences **102**(43), 15545–15550 (2005)
47. Kanehisa, M., Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research **28**(1), 27–30 (2000)
48. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Annals of Statistics, 1165–1188 (2001)
49. Collart, M.A., Panasenko, O.O.: The Ccr4–not complex. Gene **492**(1), 42–53 (2012)
50. Weskamp, K., Barmada, S.J.: RNA degradation in neurodegenerative disease. RNA Metabolism in Neurodegenerative Diseases, 103–142 (2018)
51. Cools, J.: ART: aberrant RNA degradation in T-cell leukemia. `https://cordis.europa.eu/project/rcn/185655/factsheet`. [Online] (2014-2019)
52. Dunteman, G.H.: Principal Components Analysis. A Sage Publications, vol. 69. SAGE Publications, ??? (1989)
53. Kavitha, K., Ram, A.V., Anandu, S., Karthik, S., Kailas, S., Arjun, N.: PCA-based gene selection for cancer classification. In: Procs. of the International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1–4 (2018). IEEE
54. Gniazdowski, Z.: New interpretation of principal components analysis. arXiv preprint arXiv:1711.10420 (2017)

Figures

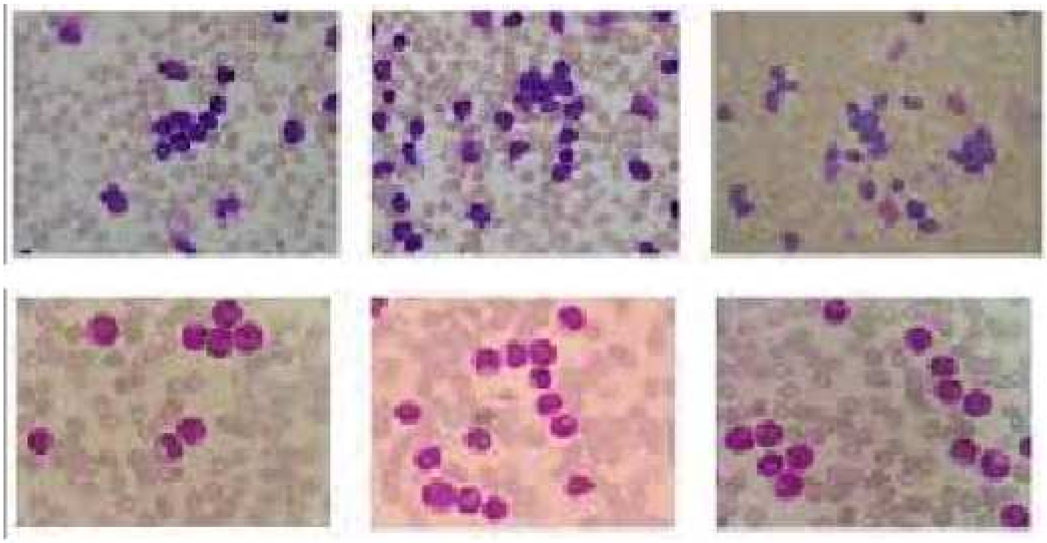


Figure 1

Acute Lymphoblastic Leukemia and Acute Myeloid Leukemia .

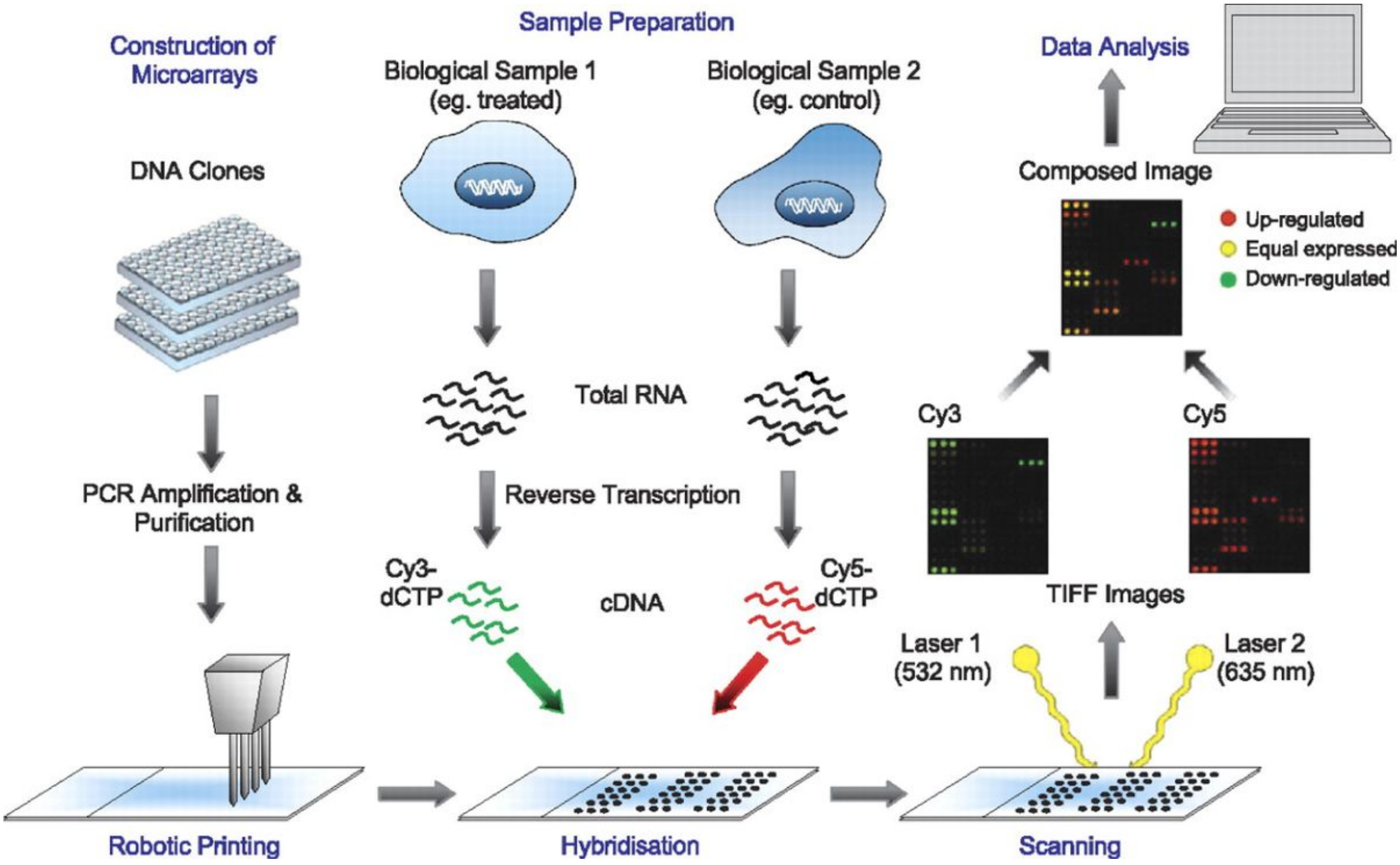


Figure 2

Visualization of the process in microarray analysis [11].

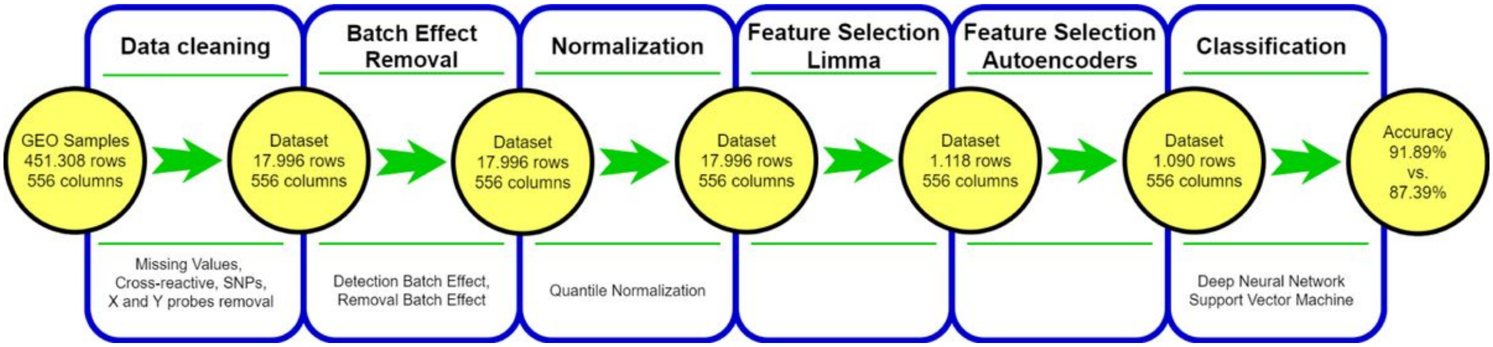

Data cleaning
Batch Effect Removal
Normalization
Feature Selection Limma
Feature Selection Autoencoders
Classification
GEO Samples 451.308 rows 556 columns
Dataset 17.996 rows 556 columns
Dataset 17.996 rows 556 columns
Dataset 17.996 rows 556 columns
Dataset 1.118 rows 556 columns
Dataset 1.090 rows 556 columns
Accuracy 91.89% vs. 87.39%
Missing Values, Cross-reactive, SNPs, X and Y probes removal
Detection Batch Effect, Removal Batch Effect
Quantile Normalization
Deep Neural Network Support Vector Machine

Figure 3

The data analysis process based on Bayesian and Autoencoders feature selection.

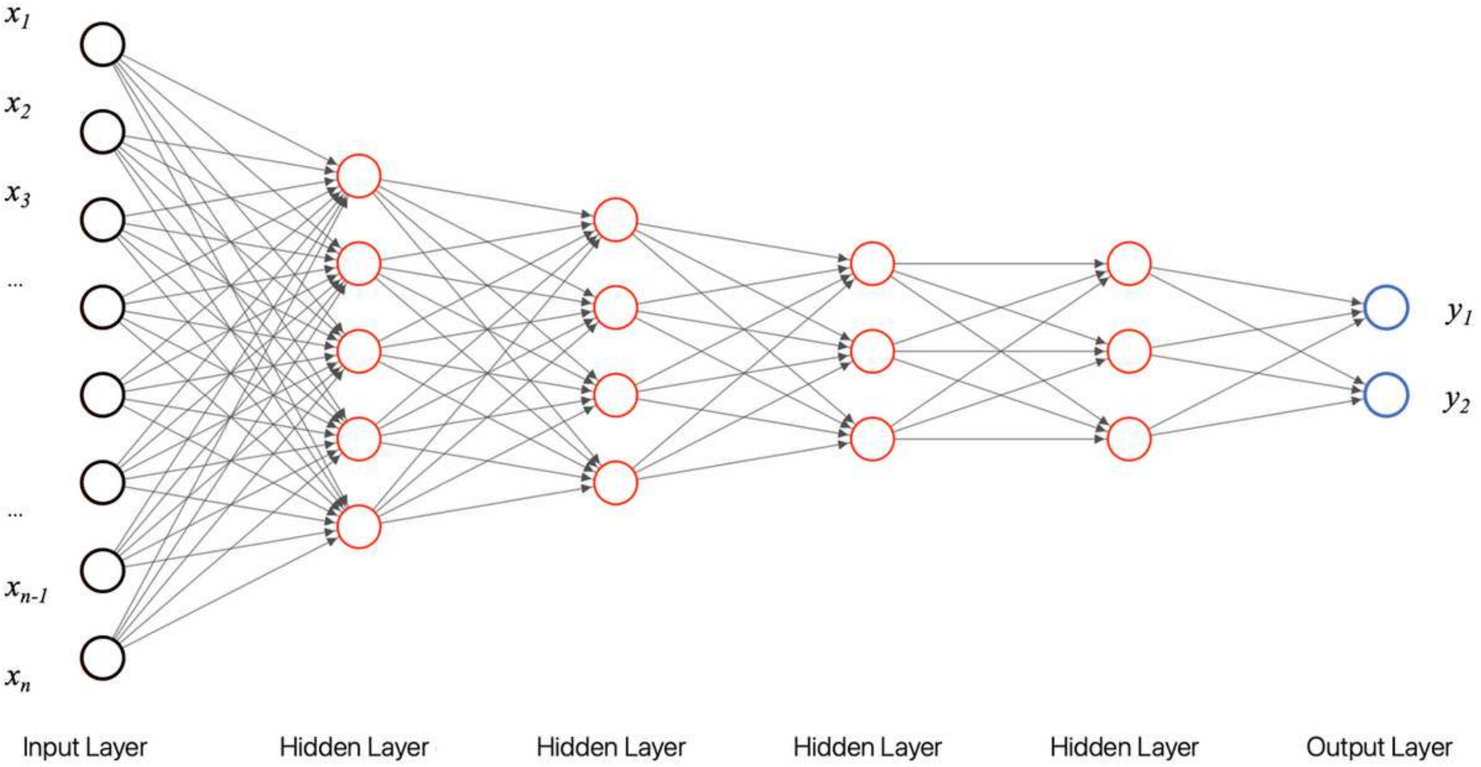


Figure 4

The DNN architecture, where n represents the number of genes.

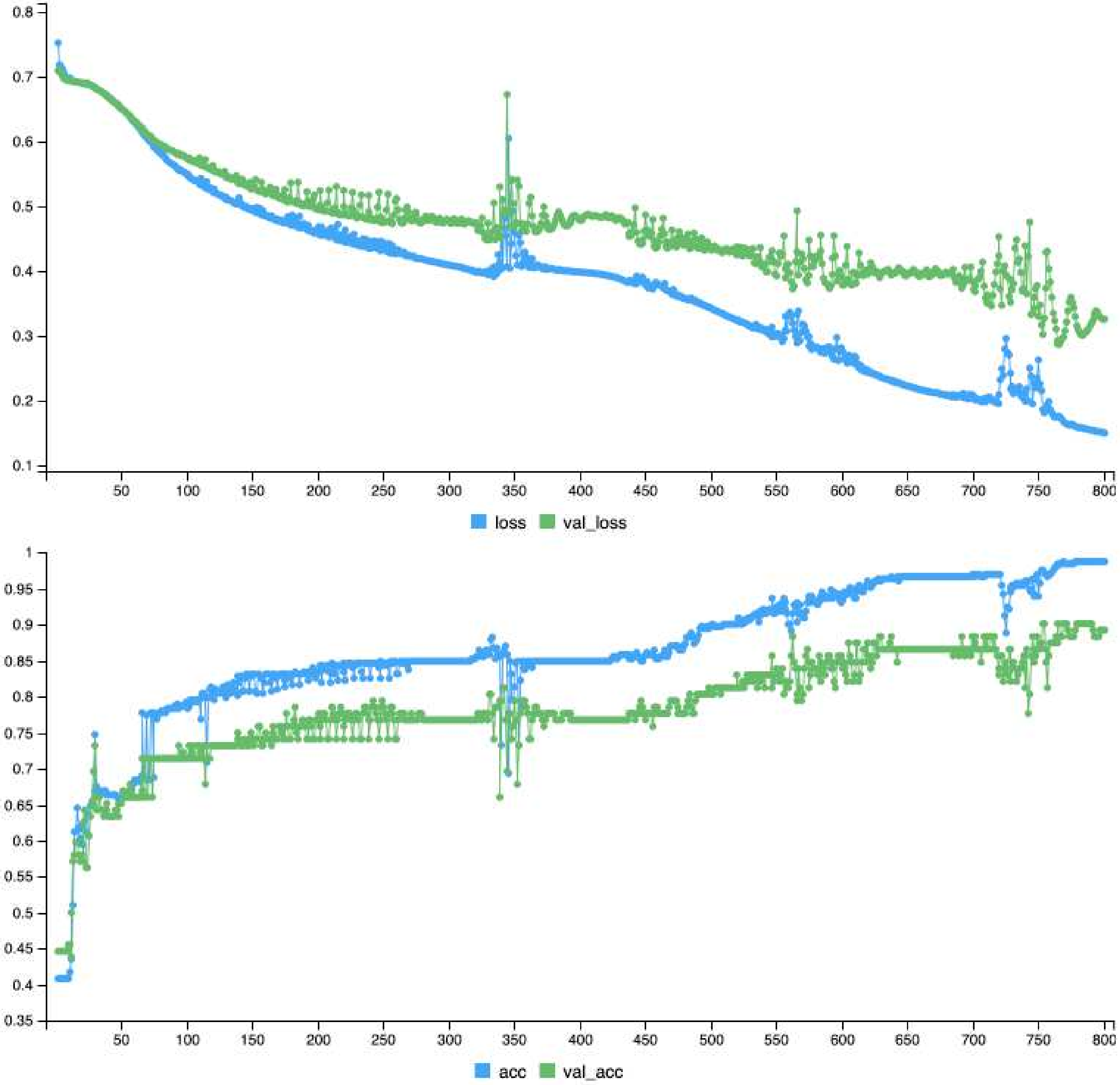


Figure 5

Loss and accuracy results of the applied neural network on the feature selection

implemented with Limma and autoencoders, where ”val”/”loss” represents the average accuracy/loss of the training set, and ”val acc”/”val loss” represents the average accuracy/loss of the validation set, respectively.

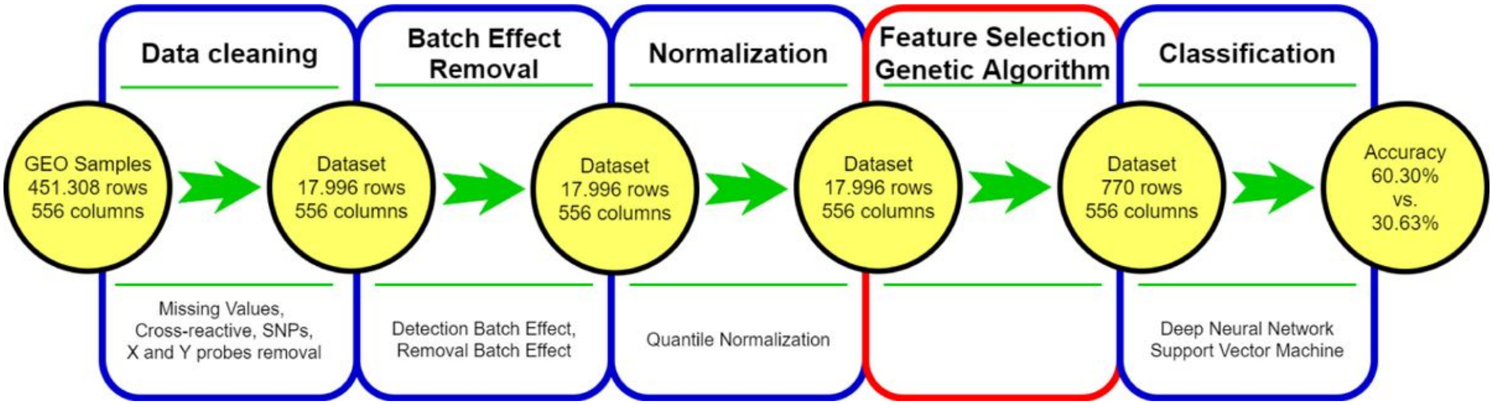

Data cleaning
Batch Effect Removal
Normalization
Feature Selection Genetic Algorithm
Classification
GEO Samples 451.308 rows 556 columns
Dataset 17.996 rows 556 columns
Dataset 17.996 rows 556 columns
Dataset 17.996 rows 556 columns
Dataset 770 rows 556 columns
Accuracy 60.30% vs. 30.63%
Missing Values, Cross-reactive, SNPs, X and Y probes removal
Detection Batch Effect, Removal Batch Effect
Quantile Normalization
Deep Neural Network Support Vector Machine

Figure 6

The data analysis process based on Genetic Algorithm feature selection.

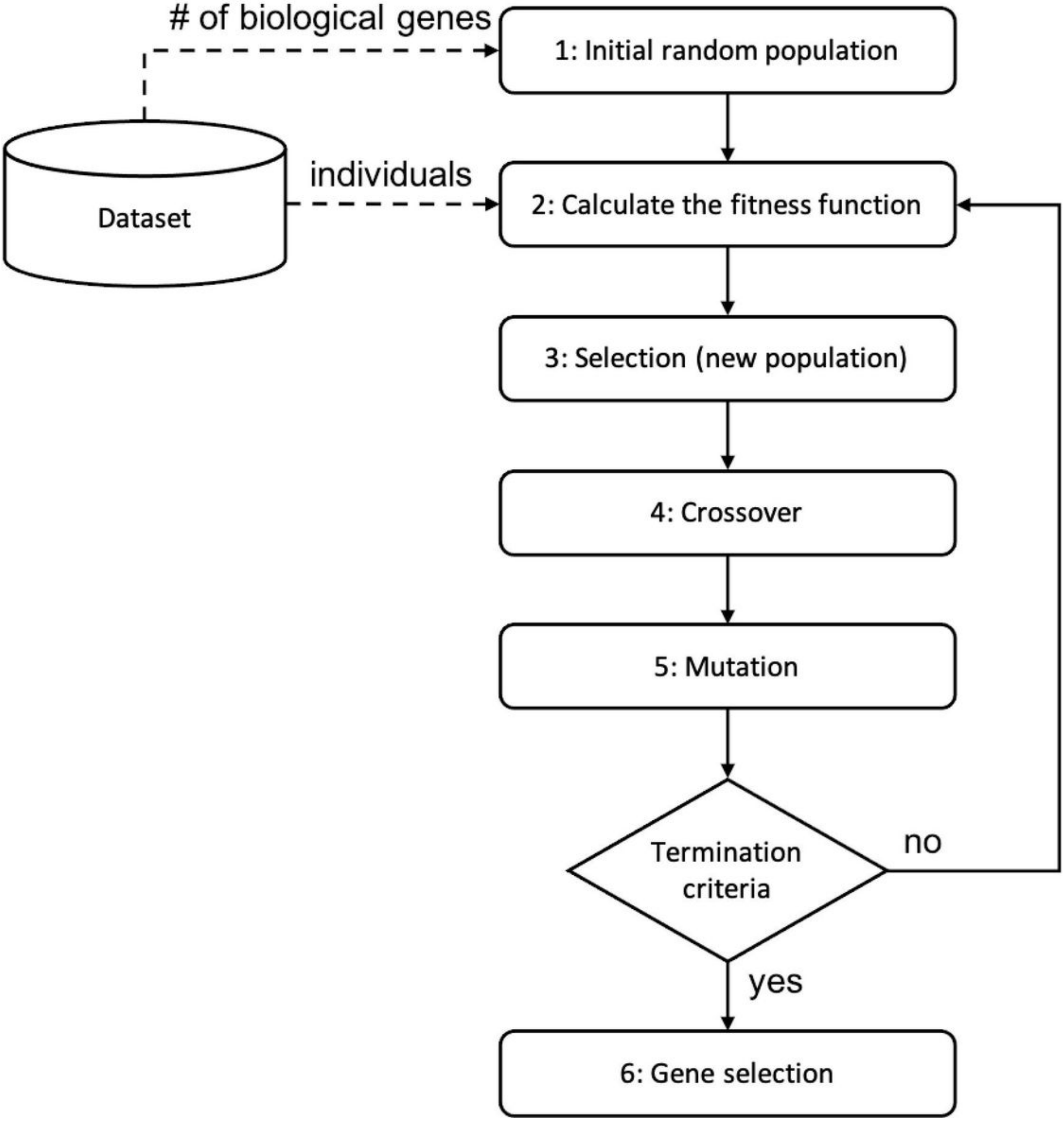


Figure 7

Steps of the GA for feature selection.

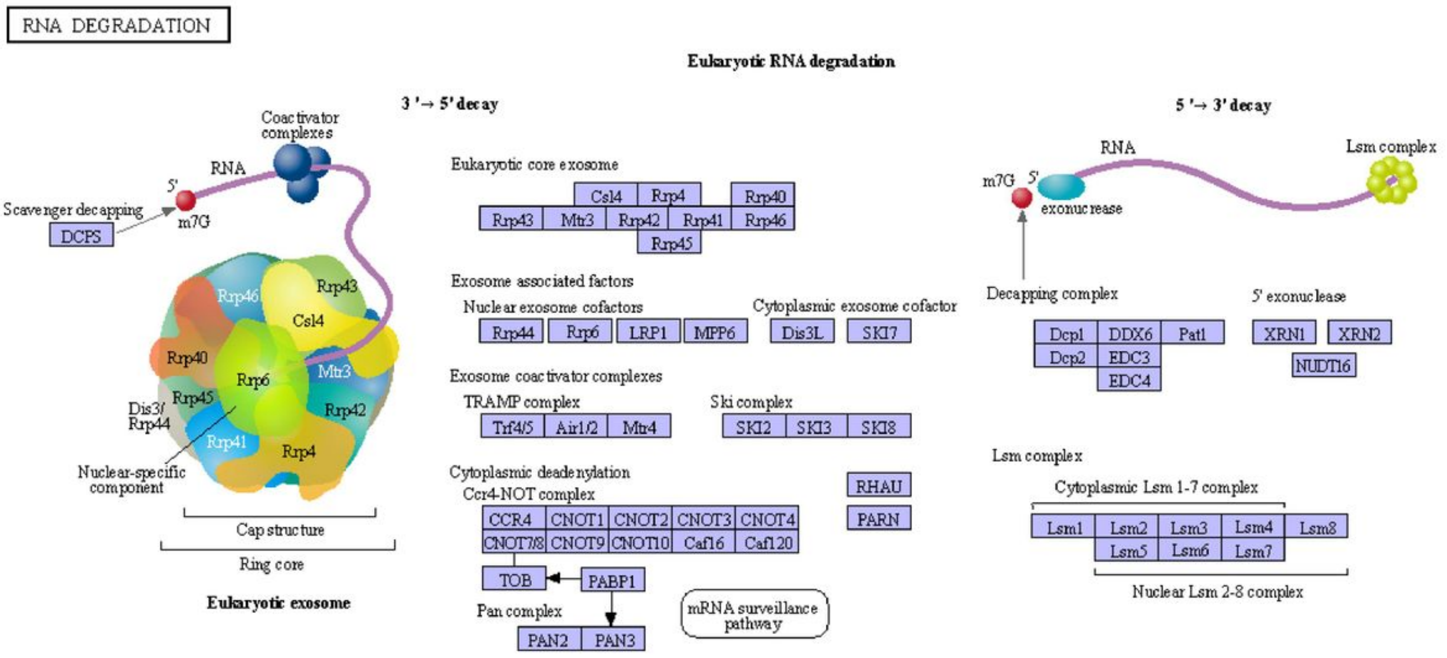


Figure 8

The ”RNA degradation” pathway [44].

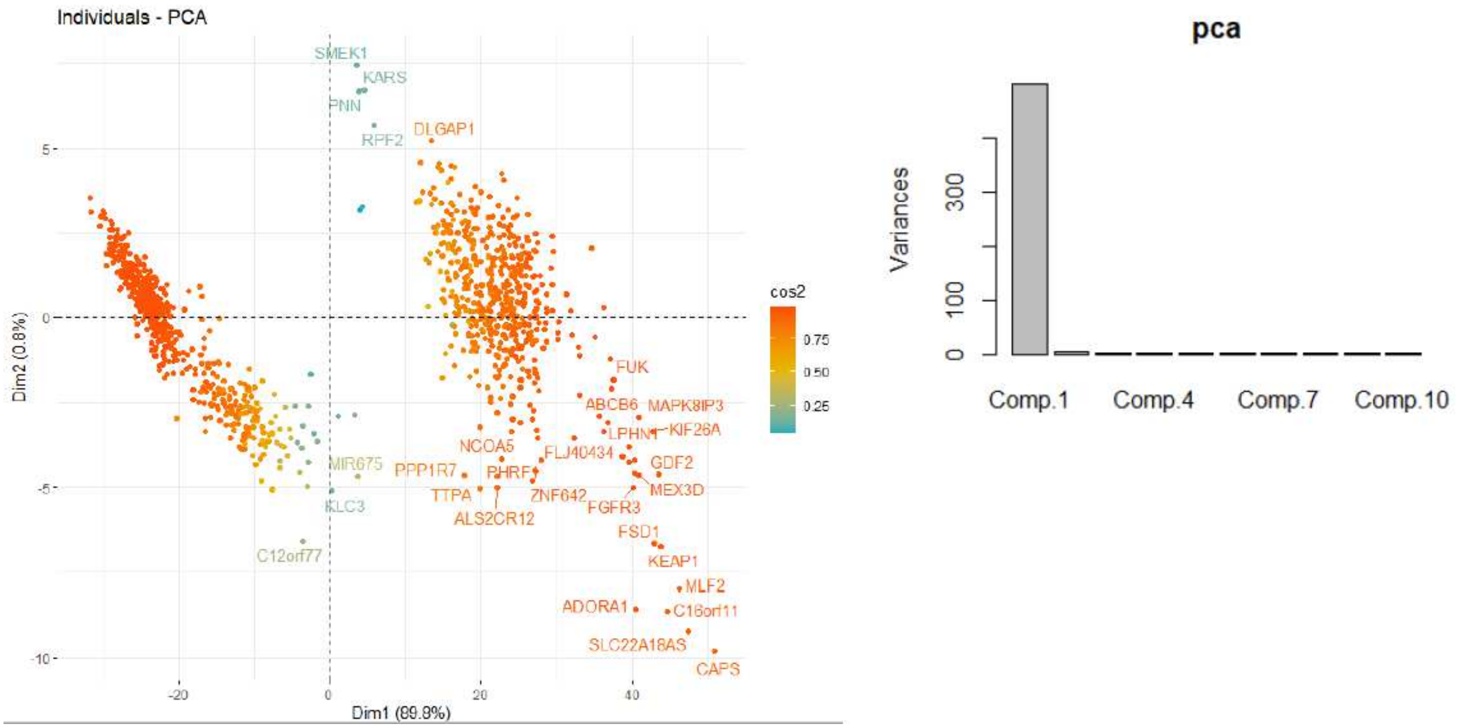
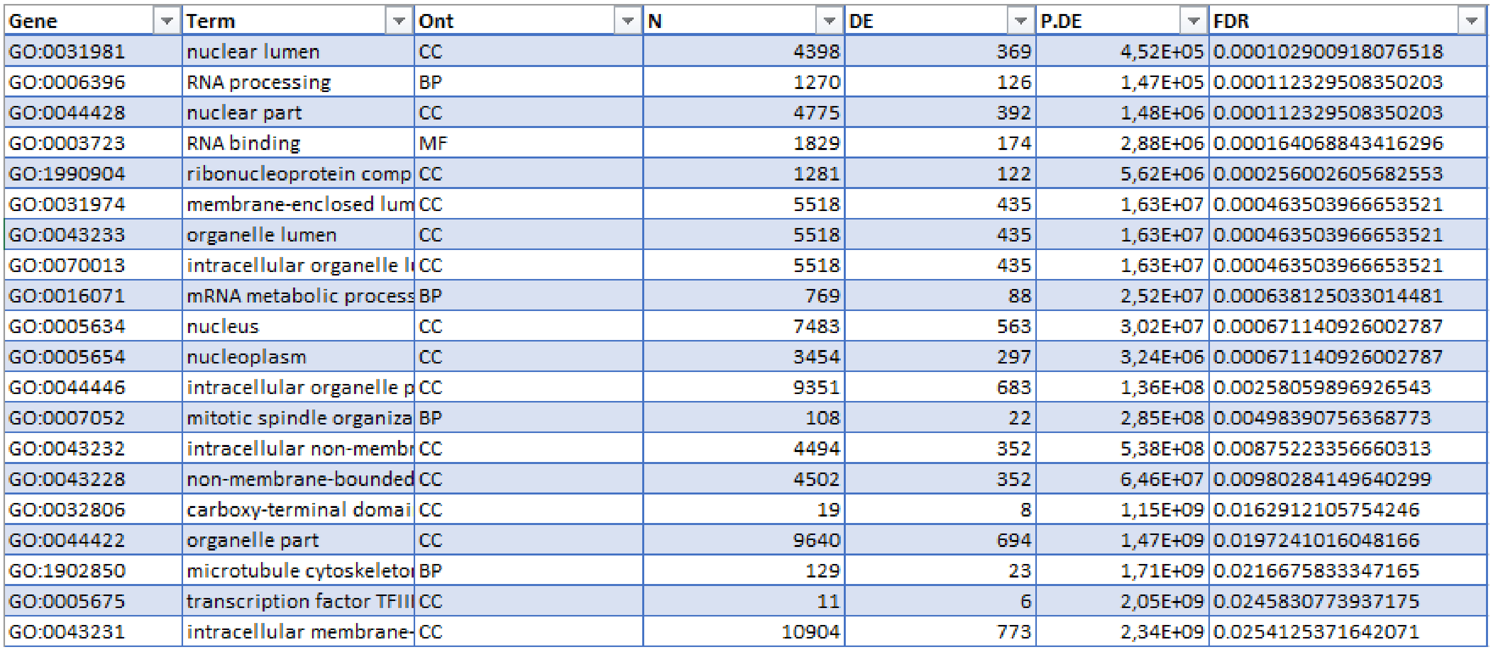


Figure 9

Results of PCA: components ?? and scores ??.



Gene	Term	Ont	N	DE	P.DE	FDR
GO:0031981	nuclear lumen	CC	4398	369	4,52E+05	0.000102900918076518
GO:0006396	RNA processing	BP	1270	126	1,47E+05	0.000112329508350203
GO:0044428	nuclear part	CC	4775	392	1,48E+06	0.000112329508350203
GO:0003723	RNA binding	MF	1829	174	2,88E+06	0.000164068843416296
GO:1990904	ribonucleoprotein comp	CC	1281	122	5,62E+06	0.000256002605682553
GO:0031974	membrane-enclosed lum	CC	5518	435	1,63E+07	0.000463503966653521
GO:0043233	organelle lumen	CC	5518	435	1,63E+07	0.000463503966653521
GO:0070013	intracellular organelle l	CC	5518	435	1,63E+07	0.000463503966653521
GO:0016071	mRNA metabolic process	BP	769	88	2,52E+07	0.000638125033014481
GO:0005634	nucleus	CC	7483	563	3,02E+07	0.000671140926002787
GO:0005654	nucleoplasm	CC	3454	297	3,24E+06	0.000671140926002787
GO:0044446	intracellular organelle p	CC	9351	683	1,36E+08	0.00258059896926543
GO:0007052	mitotic spindle organiza	BP	108	22	2,85E+08	0.00498390756368773
GO:0043232	intracellular non-membr	CC	4494	352	5,38E+08	0.00875223356660313
GO:0043228	non-membrane-bounded	CC	4502	352	6,46E+07	0.00980284149640299
GO:0032806	carboxy-terminal domai	CC	19	8	1,15E+09	0.0162912105754246
GO:0044422	organelle part	CC	9640	694	1,47E+09	0.0197241016048166
GO:1902850	microtubule cytoskeleto	BP	129	23	1,71E+09	0.0216675833347165
GO:0005675	transcription factor TFIII	CC	11	6	2,05E+09	0.0245830773937175
GO:0043231	intracellular membrane-	CC	10904	773	2,34E+09	0.0254125371642071

Figure 10

Gene enriched from ”RNA degradation” pathway.