Formation and Analysis of Gene Expression Data Based on the Joint Use of Data Mining and Machine Learning Techniques

Lyudmyla Yasinska-Damri[1], Sergii Babichev[2,3], Aleksander Spivakovsky[2] and Oleksandr Lemeshchuk[2]

[1]*Ukrainian Academy of Printing, Pid Goloskom street, 19, Lviv, 79000, Ukraine*
[2]*Kherson State University, University street, 27, Kherson, 73000, Ukraine*
[3]*Jan Evangelista Purkyne University in Usti nad Labem, Pasteurova, 15, Usti nad Labem, 400 96, Czech Republic*

Abstract

Creating a system of complex disease diagnosis based on gene expression data using modern data mining and machine learning techniques is one of the topical areas of recent bioinformatics. The main problem in this subject area consists of a large volume of experimental data; each investigated object contains approximately 20000 genes. In this paper, we propose the technique of both formation and analysis of gene expression data based on the joint use of various computational and intelligent methods of complex data processing. As the experimental data, we use the gene expression data of patients who were investigated for various types of cancer diseases. Initially, the data was formed and analyzed using the functions and modules of *TCGAbiolinks* and *Bioconductor* packages. Then, the non-informative genes in terms of both the statistical and entropy criteria were removed from the initial database. To identify the level of significance of gene expression profiles, we have applied the general Harrington desirability index, which contains, as the components, transformed statistical and entropies criteria. Finally, we applied the random forest classifier and convolutional neural network with the calculation of various types of classification quality criteria for the binary and multi-classification of the investigated objects in the first and the second cases, respectively. Analyzing the simulation results has shown that the identification accuracy of the examined samples is high in all cases. To our mind, the proposed technique creates the basis for improving complex disease diagnosis systems based on gene expression data.

Keywords 1

Gene expression profiles, Shannon entropy, statistical criteria, Harrington desirability function, Random Forest (RF), Convolution Neural Network (CNN), classification quality criteria

1. Introduction

Modern artificial information processing systems used in various fields of both intelligent data analysis and machine learning are, in most cases, based on the analogy of the functioning of relevant processes in biological organisms. Such processes should include the functioning of the gene network, immune processes, functioning of neural networks, etc. A particularity of such systems is a high level of complexity, the ability to learn, parallelization of information processing, a high level of protection, and the ability to recognize and make adequate decisions. In this context, developing modern artificial models of big data processing is possible using a systematic approach, which involves combining knowledge and methods from various practical fields, such as molecular biology, mathematics, computer science, physics, and chemistry. This approach creates conditions for increasing the

IntelITSIS'2023: 4th International Workshop on Intelligent Information Technologies and Systems of Information Security, March 22–24, 2023, Khmelnytskyi, Ukraine
EMAIL: Lm.yasinska@gmail.com (L. Yasinska-Damri); sergii.babichev@ujep.cz (S. Babichev); spivakovsky@ksu.kherson.ua (A. Spivakovsky); olemeshchuk@ksu.ks.ua (O. Lemeshchuk)
ORCID: 0000-0002-8629-8658 (L. Yasinska-Damri); 0000-0001-6797-1467 (S. Babichev); 0000-0001-7574-4133 (A. Spivakovsky); 0000-0002-9876-3502 (O. Lemeshchuk)
© 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3373/paper1.pdf

objectivity of processing big data in real-time due to the use of ensembles of methods, hybrid models and effective quality criteria for evaluating results at the appropriate stage of the data processing.

One of the actual problems of modern bioinformatics is the processing of gene expression data obtained by performing DNA microarray experiments or by the more modern method of RNA molecule sequencing. The main particularity of the experimental data is the large number of attributes that define the state of the corresponding biological organism. As well know, the number of expressed (active) genes that make up the human genome is approximately equal to 25,000. At the same time, the number of each type of gene varies within the range from 0 to hundreds of thousands ones. Creating systems of complex objects diagnosis or models of gene network reconstruction based on complete data is ineffective since interpreting the obtained results is problematic in this instance. Deep learning models such as deep neural networks, convolutional neural networks, and deep artificial networks exist and they are actively used for processing big data, allowing for obtaining satisfactory diagnostic accuracy on big data. Still, at the same time, there is a problem with training time, network sensitivity, and verification received models on other similar data. Solving the abovementioned problem is possible by developing new and improving existing intelligent models, information technologies and algorithms focused on big data processing. The implementation of this process assumes the formation of high-quality experimental data by applying adequate methods of gene expression values normalization and reduction of lowly-expressed genes at the first stage. In the second stage, the problem of forming subsets of differentially expressed and mutually correlated gene expression profiles arises by applying hybrid models of high-dimensional data clustering. The third stage is developing a medical diagnosis model on the basis of the allocated clusters of expression values using deep learning methods with appropriate verification of the obtained results.

However, we would like to note that in spite of specific achievements in this field, the effective solution to the problem of gene expression data processing in order to create a medical diagnostic system that allows us to identify with acceptable accuracy the patient's state at an early stage of the disease is absent nowadays. To our mind, increasing the effectiveness of current intelligent systems focused on gene expression data processing for the following creation of various disease diagnostic systems can be achieved by providing the following:

- justified choice of methods for formation and analysis of experimental data to form an array of gene expression values for the investigated objects;
- application of effective metrics for assessing the proximity between gene expression profiles based on the assessment of mutual information between the respective profiles;
- justified shaping subsets of diversely expressed and jointly correlated gene expression profiles using effective and adequate clustering quality criteria;
- the creation of hybrid models for diagnosing the state of the investigated object using inductive models of objective clustering and deep neural networks;
- justified choice of methods to validate the models of complex disease diagnosis with the application of appropriate quality criteria for evaluating the model's effectiveness on the test subset of expression values.

The **goal of the paper** is to develop the model of formation, analysis and processing of gene expression data on the basis of the use of the functions of both *TCGAbiolinks* and *Bioconductor* packages of *R* software, Random Forest classification algorithm and Convolutional Neural Network.

2. Related works

Recently, much scientific research focused on solving the challenges of formation and analyzing gene expression data for their reduction and further application in models for diagnosing the object state or the gene regulatory network reconstruction. So, the paper [1] proposed a hybrid method for extracting genes by the joint use of the Information Gain method and the Standard Genetic Algorithm. This method was named as IG/SGA method. Initially, the IG information increment method was used to reduce the non-informative objects. Then, the data was processed using a genetic algorithm. Finally, the authors have applied the genetic programming (GP) method to carry out the object classification procedure. The simulation procedure assumed the use of seven sets of gene expression data of patients investigated for cancer disease. The authors have shown that the application of the offered technique can allow them

to reach 100% classification correctness in two instances. In other cases, the classification accuracy was less.

In [2,3], the authors showed preferences of the Ant Colony Optimization (ACO) algorithm application when the hybrid models of complex data processing were developed. So, the method offered in [2] is based on the joint application of cellular learning automata and the ACO method (CLA-ACO). The practical realization of the proposed pattern assumes three phases. Initially, the Fisher-filtered method is applied to the data. At this phase, the non-informative genes in expression values were removed from the initial data. Then, CLA and ACA techniques with optimal functions are applied to a subset of gene expression profiles. In the third phase, the authors formed and analyzed the final subsets of gene expression data by assessing the errors of both the first and the second types using the ROC analyzing method. The presented in this manuscript testing results showed the high performance of the offered method for the classification of the examined objects. In [3], the authors combined ACO algorithm and Adaptive Stem Cell Optimization (ASCO) technique for the purpose of forming gene expression datasets. The proposed model filtered the data by assessing the mutual information values in the first phase. Then, various types of classifiers were applied to the formed datasets to assess the model performance. Analyzing the received results, in this instance, also indicates the high performance of this pattern.

An analysis of various current hybrid patterns focused on forming the subsets of gene expression datasets for the purpose of creating or improving diagnostic systems based on gene expression data is presented in [4]. The authors analyzed in detail, with the allocation of advantages and shortcomings, various intelligent systems and algorithms such as the Mutual Information Maximization (MIM) filtering method, Genetic Algorithm (GA) grouping method and SVM classifier [5]; Correlation-based Feature Selection (CFS) filtering method, Genetic Algorithm (GA) grouping method and KNN classifier [6]; Laplacian and Fisher score filtering method, Genetic Algorithm (GA) grouping method and SVM, KNN and NB classifiers [7]; Independent Component Analysis (ICA) filtering method, Artificial Bee Colony (ABC) grouping method and NB classifier [8] and other hybrid techniques based on the combined application of various data mining and machine learning algorithms [9-14].

The analysis of various hybrid model operation results has allowed the authors to remark that the challenger of objective formation of subsets of genes, which can allow us to recognize the examined samples with high-reliability levels, does not unambiguous solved nowadays. In most proposed patterns, high classification performance is ripened when applying a small number of genes. This problem can be solved by using deep learning methods at the phase of gene expression data classification. So, in [15] the authors describe the technique focused on the early diagnosis of cancer disease using a Deep Neural Network (DNN). The proposed model was tested on 37 different kinds of cancer. The authors used datasets from the TCGA cancer genome atlas. The experimental dataset consisted of 10,663 samples (9,807 tumors and 856 normal samples) for 37 cancer types, and they contained the expression of 10,000 genes. The models of deep neural networks with three different structures were investigated: three-levels (3NN), five-levels (5NN) and nine-levels (9NN), while a comparative analysis was performed with the model on the basis of the use of SVM method by assessing the classification performance. An analysis of the obtained results has shown that the highest diagnostic accuracy was achieved when using the 5NN model, but the performance of disease classification was not satisfactory. Moreover, the sensitivity of the model to the type of cancer is also not high. This fact indicates that the proposed model needs to be refined.

In [16], the authors investigated different types of patterns based on the use of an ensemble of DM and ML algorithms, including deep learning methods for the diagnosis of three types of cancer: lung, breast, and stomach cancer. 162 lung cancer samples, 878 breast cancer samples, and 271 stomach cancer samples were used. The accuracy of disease diagnosis for all data sets was 98%. However, it should be noted that despite the impressive results, the proposed model has several significant shortcomings. First of all, the proposed model is very complex and requires both a lot of time on model training and large computer resources. The second significant drawback is that the model based on an ensemble of deep learning methods is challenging to interpret.

In [17], the authors investigated several structures of Convolutional Neural Networks (CNN) with various combinations of hyper-parameters for the diagnosis of different types of cancer using unstructured gene expression data. During the simulation process, three models of CNN with different combinations of hyper-parameters were investigated: one-dimensional neural network model 1D-CNN;

two-dimensional CNN 2D-Vanilla-CNN; hybrid convolutional neural network 2D-Hybrid-CNN (as input contains a two-dimensional matrix and a one-dimensional vector of gene expressions). The models were trained and tested on gene expression data containing 10,340 samples from 33 types of cancer and 713 samples taken from normal tissues (no tumor detected). The data contained 7100 genes whose expression was used as attributes. Data were also taken from the Cancer Genome Atlas (TCGA). Analyzing the authors' simulation results showed the high performance of the proposed patterns. For the diagnosis of 34 classes (33 with cancer tumors and one normal), the prediction accuracy on the test subset of data varied within the range from 93.9% to 95%. However, it should be noted that training the proposed models also requires much time. In addition, the authors used unstructured data, which may also affect the accuracy and sensitivity of the proposed models operating.

The hereinbefore presented brief analysis of the current state of research regarding the formation, analysis and processing of expression data to develop and improve disease diagnostic systems allows us to conclude that nowadays, there is no effective and complete information technology for processing gene expression data to diagnose complex diseases. Existing models of disease diagnosis systems on the basis of gene expression data can allow us to identify the relevant disease with a certain accuracy. However, they have certain shortcomings, which are associated with the imperfection of big data processing models for selecting informative gene expression profiles, low objectivity in the formation and further processing of gene expression profiles, low objectivity of decision-making regarding the presence or absence of the relevant disease, etc. These facts indicate the actuality of the presented research.

3. Gene expression data formation and analysis

We used gene expression data of patients examined for various cancer kinds from The Cancer Genome Atlas (TCGA) [18]. The formation of experimental data was carried out using the functions of the *TCGAbiolinks* package [19], which is a part of the *Bioconductor* package [20] of the *R* software [21]. The preference of this data is determined by the fact that cancer is currently one of the primary reasons for death globally, and treatment methods for this disease vary in a wide range. TCGA database, which was launched in 2006 to collect and analyze different data for more than 35 various kinds of cancer by selecting many cases of different tumor types, nowadays is one of the most comprehensive human cancer databases, containing various kinds of data. Tumors formed using TCGA tools varied from solid to hematologic, from mild to highly aggressive considering survival, and from beginning to metastatic. Thus, the experimental gene expression data contained in TCGA is a very powerful platform for both the approbation and evaluation of the effectiveness of the appropriate model focused on processing gene expression data in order to create a hybrid model for the diagnosis of complex diseases. The advantages of the TCGAbiolinks package should also include the fact that its use can allow us to download and generate data from various platforms that are used in current times in various research institutions to generate experimental data based on genome analysis. Moreover, the functions of this package can allow us to implement various types of data formation, analysis and processing using various patterns of visualization of the particularities of data distribution in the relevant features space.

Within the framework of the research, we used gene expression data obtained by applying the method of RNA molecule sequencing received on the Illumina platform [22]. Each sample of initial data contained, as attributes, 19947 types of genes, while the expression of the corresponding gene is proportional to its activity level. The experimental database contained nine types of data, eight of which corresponded to eight types of cancer, and the ninth data corresponded to samples for which no cancer was identified as a result of clinical trials. The total number of studied samples was equal to 3269. Thus, at the initial stage of modelling, the gene expression matrix had the following form: $E = (3296 \times 19947)$. The classification of experimental data is presented in Table 1 and Figure 1.

Table 2 shows the general statistics of the number of genes' maximum values distribution, which contained experimental data for all the studied samples. Analyzing the data in Table 2 indicates that there is a certain number of non-expressed genes for all samples (zero number) and a certain number of low-expressed genes (the number of active genes for all samples is significantly less than the number of other genes). These genes can be removed from the data without significant loss of useful information.

Table 1
Experimental datasets of gene expression data

No	Kind of cancer	Number of samples
1	Adrenocortical carcinoma - ACC	79
2	Glioblastoma multiforme - GBM	169
3	Sarcoma - SARC	263
4	Lung squamous cell carcinoma - LUSC	502
5	Lung adenocarcinoma - LUAD	541
6	Stomach adenocarcinoma - STAD	415
7	Kidney renal clear cell carcinoma - KIRC	542
8	Brain Lower Grade Glioma - LGG	534
9	Normal (without tumor)	224

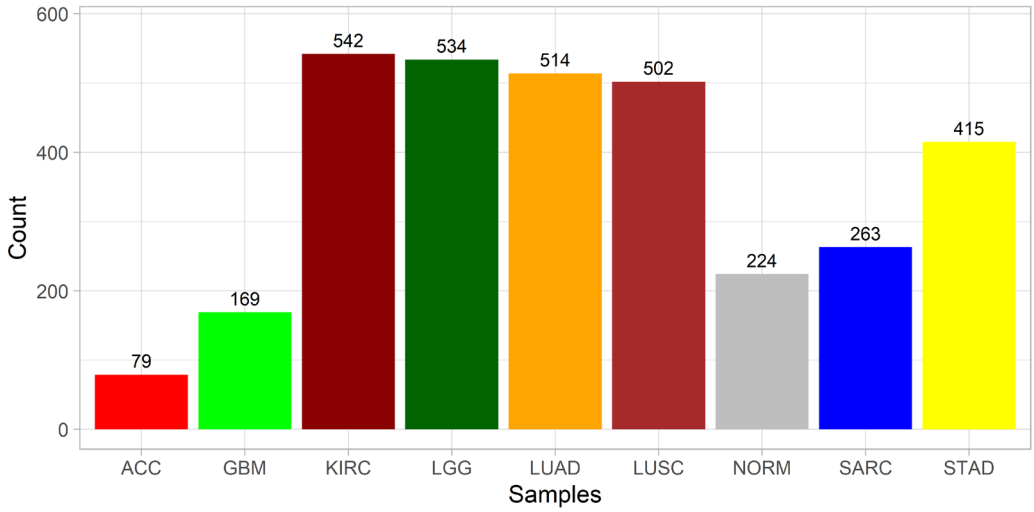

Figure 1: The bar chart of the sample's distribution in the gene expressions experimental data for the various kinds of cancer

Table 2
General statistics of the number of genes' maximum values distribution for all samples

Min	1st quartile	Median	Mean	3rd quartile	Max
0	3775	12030	44324	30945	6303246

Moreover, the analysis of the data in Table 2 also shows the need to transform the absolute values of the number of expressed genes into more convenient values with a smaller range of their variation. The Bioconductor package [20] offers a method of transforming the absolute values of the number of genes into a relative value (count-per-million - *cpm*), which can be calculated as follows:

$$count'_{ij} = \frac{count_{ij}}{\sum_{j=1}^{m} count_{ij}} \cdot 10^6 \quad (1)$$

where: $count_{ij}$ is the number of the j-th gene for the i-th sample; m is the number of the examined genes.

The general statistics of the transformed by formula (1) the number of genes maximum values distribution for all investigated samples are presented in Table 3.

Table 3
General statistics of the transformed quantity of genes' maximum values distribution for all samples

Min	1st quartile	Median	Mean	3rd quartile	Max
0	76.73	237.87	886.58	598.56	172693.66

As can be seen, the variation range of the transformed values is significantly smaller than the initial data presented in Table 2, which simplifies the procedure for further processing of experimental data. The next step of the data preparation is the removal from the experimental data of unexpressed gene expression profiles by the corresponding thresholding coefficient. In our experiment, the value of this coefficient corresponds to the condition $log_2(count'_{ij}) = 0$ for all investigated samples. According to this condition, genes whose profiles correspond to the condition $\max(count') \leq 1$ are considered non-expressed, and they are removed from the data. At this stage, the number of genes was reduced by 682, and the gene expression matrix took the following form: $E = (3296 \times 19265)$.

At the next stage, the values of gene expressions were transformed into a more convenient range by applying the function $log_2(\cdot)$ to all values of the matrix. In this instance, the simulation results have shown that some values smaller than one were transformed into corresponding negative values. This is unacceptable because the minimum gene expression value is 0 (the gene is inactive). Therefore, at this stage, negative values were replaced by zero, corresponding to non-expressed genes. Figure 2 shows the distribution of normalized expression values of nine genes for one adrenocortical carcinoma (ACC) sample. Figure. 3 shows the nature of the expression values distribution for one type of gene for all investigated samples. In this case, the genes A1BG.1 (Fig. 3a), A2M.2 (Fig. 3b), NAT1.9 (Fig. 3c) and NAT2.10 (Fig. 3d) were studied. The analysis of the obtained results allows us to conclude that the nature of the gene expression values distribution differs significantly for different samples both in absolute value and variance. This fact creates the conditions for identifying samples by expression values when applying the appropriate classifier.

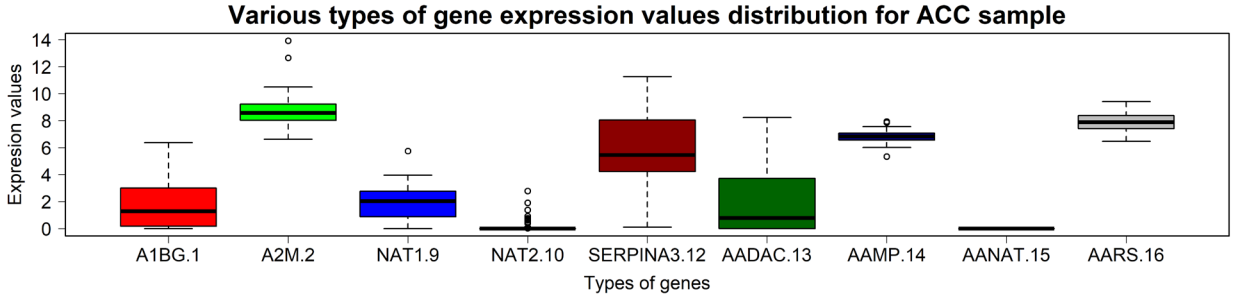

Figure 2: Boxplots of the expression values distribution of different types of genes for one sample of adrenocortical carcinoma (ACC)

The next phase of the experimental data processing is the extraction of uninformative genes considering various kinds of statistical and entropy criteria using Harrington's desirability function by the technique described in detail in [23].

4. Filtering gene expression profiles by statistical and entropy criteria

As statistical criteria, we used the maximum absolute value of the expression assessed for all samples and the variance of the appropriate gene expression profile. Shannon entropy was calculated using the James-Stein shrinkage estimator. Figure 4 shows the boxplots of the distribution of the various criteria values. Analyzing the charts indicates an increase in the gene expression profile significance with an increase in the statistical criteria values and a decrease in the value of the entropy criterion. In accordance with the above, the algorithm for calculating a complex criterion based on the Harrington method, which determines the level of the significance of the genes, assumes the following phases:

1. Assessing the linear equations coefficient values. At this phase, we used the boundary values of both the appropriate criteria and Y dimensionless parameter ($Y_{min} = -2$; $Y_{max} = 5$):

$$\begin{aligned} Y_{min}^{(mabs,var)} &= a^{(ambs,var)} + b^{(mabs,var)} \cdot \min(mabs, var) \\ Y_{max}^{(mabs,var)} &= a^{(mabs,var)} + b^{(mabs,var)} \cdot \max(mabs, var) \\ Y_{min}^{(entr)} &= a^{(entr)} - b^{(entr)} \cdot \max(entr) \\ Y_{max}^{(entr)} &= a^{(entr)} - b^{(entr)} \cdot \min(entr) \end{aligned} \quad (2)$$

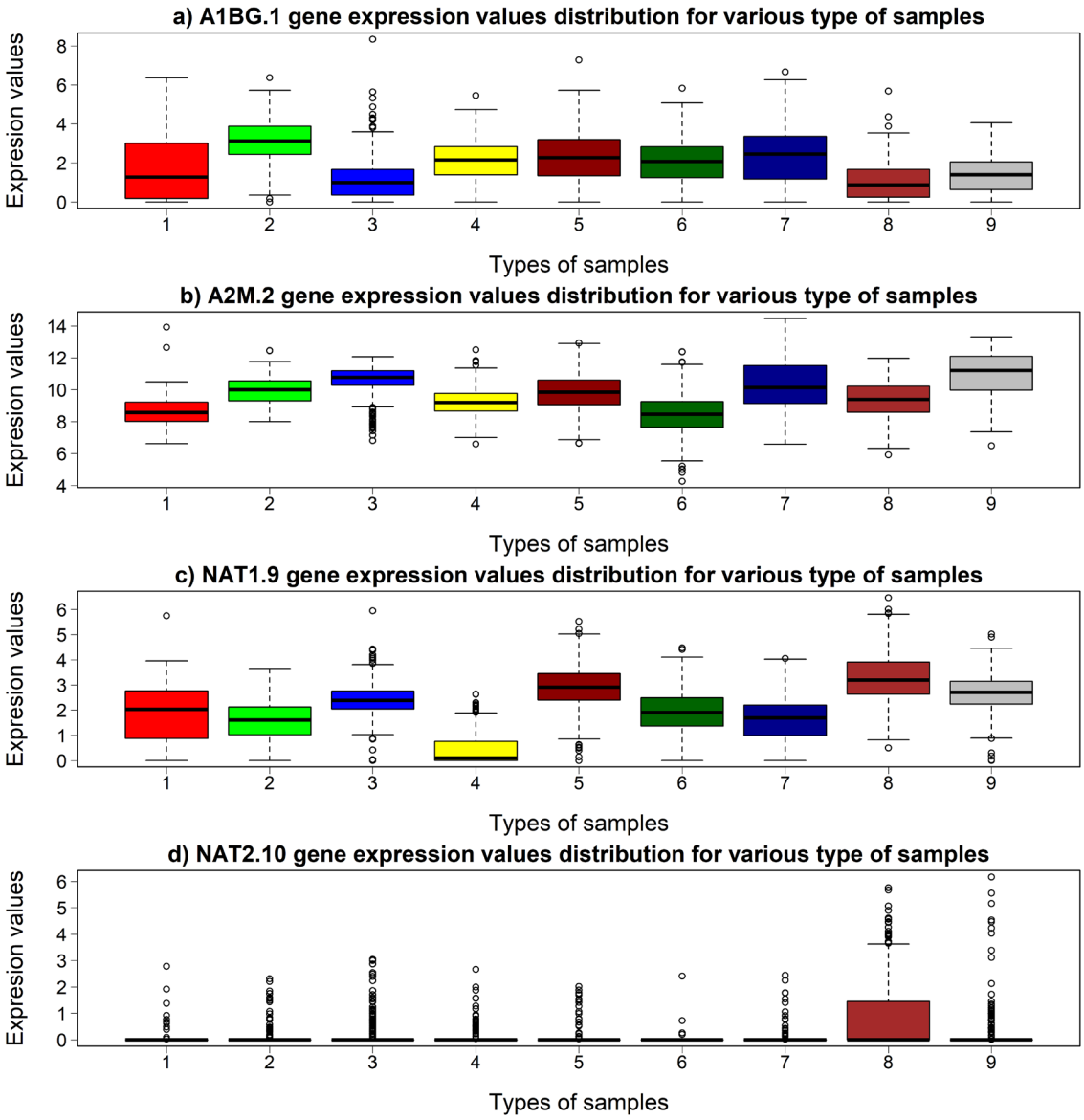

Figure 3: Boxplots of the expression values distribution of one type of gene for different samples. The studied genes were: a) A1BG.1; b) A2M.2; c) NAT1.9; d) NAT2.10

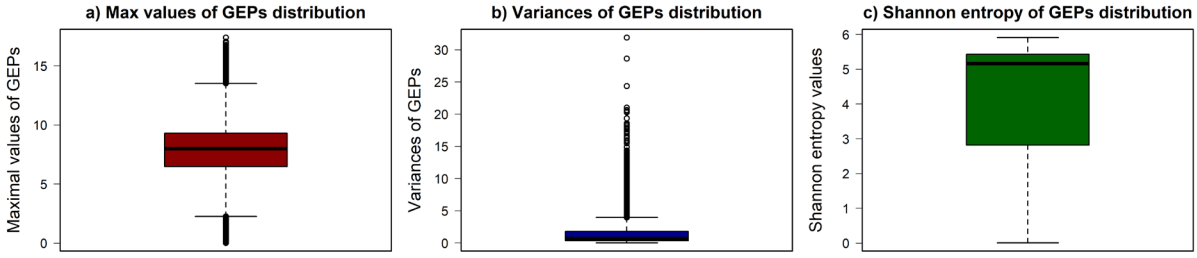

Figure 4: Boxplots of the distribution of the gene expression profiles statistical criteria values and Shannon's entropy: a) maximum expression value; b) variance; c) Shannon entropy

2. Calculation of the Y parameter values for each value of the criteria:

$$\begin{aligned} Ym^{(abs)} &= a^{(mabs)} + b^{(mabs)} \cdot \max\_abs \\ Y^{(var)} &= a^{(var)} + b^{(var)} \cdot \mathrm{var} \\ Y^{(entr)} &= a^{(entr)} - b^{(entr)} \cdot \mathrm{entr} \end{aligned} \tag{3}$$

where: *max_abs* is the maximum absolute expression values of the corresponding gene profile; *var* and *entr* are the Shannon entropy end variance of this gene expression profile, respectively;

3. Evaluation of d-values for each value of the Y parameter:

$$\begin{aligned} d^{(mabs)} &= \exp\left(-\exp\left(-Y^{(mabs)}\right)\right) \\ d^{(var)} &= \exp\left(-\exp\left(-Y^{(var)}\right)\right) \\ d^{(entr)} &= \exp\left(-\exp\left(-Y^{(entr)}\right)\right) \end{aligned} \tag{4}$$

4. Evaluation of the generalized index D as a geometric average of all d-values. This index indicates the gene expression profile significance:

$$D^{(general)} = \sqrt[3]{d^{(mabs)} \cdot d^{(var)} \cdot d^{(entr)}} \tag{5}$$

A higher value of the index (5) corresponds to a higher significance of the gene expression profile according to the used criteria. Table 4 presents the general statistics of the general index values distribution.

Table 4

General statistics of the general index values distribution

Min	1st quartile	Median	Mean	3rd quartile	Max
0.007239	0.029438	0.039786	0.074478	0.075635	0.947162

Filtering gene expression profiles at this stage involved removing profiles with a generalized desirability index lower than the value of the index corresponding to the first quartile, i.e., 0.029438. The number of gene expression profiles, in this case, was reduced by 4814, i.e., the filtered matrix of gene expression values took the form: $E = (3296 \times 14451)$.

5. Evaluation of the proposed technique adequacy using the machine learning techniques

Assessment of the gene expression profiles filtering model adequacy was performed in two stages. In the first stage, a binary Random Forest classifier was applied to the samples, including the complete set of gene expressions (19265) and the set of gene expressions after removing non-informative ones by the used criteria (14451). The efficiency of using the Random Forest classifier for the binary identification of samples used as the attributes of big data is proved in [24]. In the first step, the complete set of samples was divided into eight subsets, each of which contained gene expressions of samples for which no tumor was detected (224 samples) and samples with a cancerous tumor, taking into account the type of disease. In accordance with the methodology presented in [24,25], the experimental data were divided into two subsets, 65% of the samples were used for training the classifier, and 35% for testing the obtained model with the evaluation of data classification performance.

In the second phase, we applied the convolutional neural network (CNN) for complete and filtered gene expression data to solve a multi-class task. Taking into account the simulation results presented in [26], we used a one-dimensional two-layered CNN, the structural topology of which is shown in Figure 5. For the correct initialization of the first and second layer filters, the experimental data were supplemented with profiles with zero expression up to 19300 and 14500 in the case of using complete and filtered gene expression data, respectively. The kernel size was set to 8, and the compact layer density was 256. The results were evaluated on a test subset of the data, the number of objects of which was 35% of the total number of investigated objects (1144 out of 3269).

Table 5 shows the simulation results regarding the binary classification of the investigated samples using the complete set of gene expression profiles (19265 genes). It should be noted that when applying filtered data (14451 genes), the results were almost the same. Analysis of the data in Table 5 indicates the high accuracy of binary classification using complete gene expression data. In three cases, the model identified the patient's state with 100% accuracy. In the other five cases, the accuracy of identifying the patient's state according to the appropriate classification criteria is not 100% but is also very high.

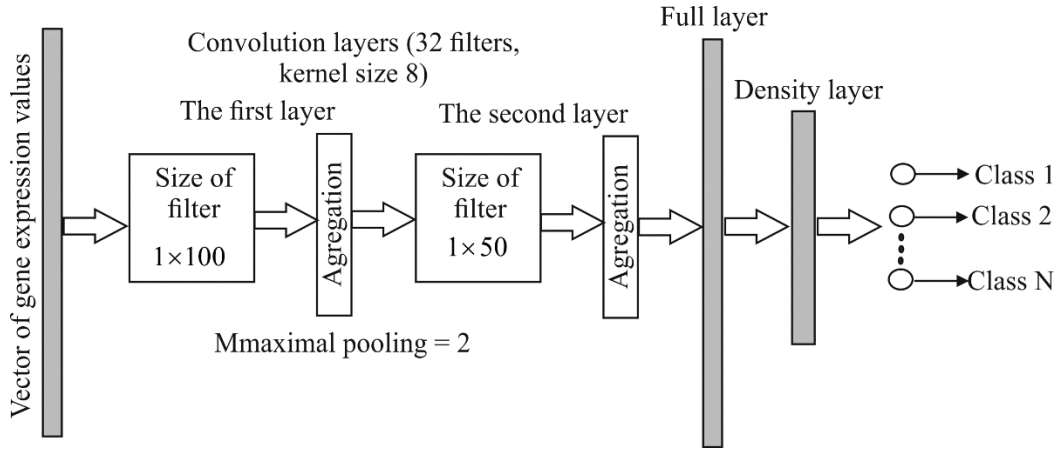

Figure 5: Topology of one-dimensional 2NN CNN used as multi-classifier to identify both the state of the patient and type of cancer

Table 5

Results of the investigated samples binary classification using the complete set of gene expression profiles

Type of data	Classification quality criteria				
	Accuracy	Sensitivity	Specificity	F-measure	MCC
acc	1	1	1	1	1
gbm	0.993	0.983	1	0.988	0.985
kirc	1	1	1	1	1
lgg	0.992	0.995	0.987	0.994	0.982
luad	0.989	0.984	1	0.987	0.973
lusc	0.992	0.989	1	0.990	0.982
sarc	1	1	1	1	1
stad	0.987	0.963	1	0.975	0.971

Tables 6 and 7 present the research results regarding applying a CNN to solve a multiclass task based on complete and filtered gene expression data, respectively.

Table 6

The results of the application of CNN for the classification of various types of cancer when using complete data (19265 genes)

Class	Number of samples	Sensitivity	Specificity	F-measure	Accuracy
ACC	24	0.85	0.96	0.9	
GBM	73	0.97	0.93	0.95	
KIRC	169	0.99	0.98	0.99	
LGG	191	0.97	0.98	0.98	97%
LUAD	200	0.96	0.96	0.96	
LUSC	169	0.96	0.96	0.96	
SARC	73	0.96	0.96	0.96	
STAD	96	0.94	0.96	0.95	
NORMAL	149	0.99	0.98	0.99	

The simulation results indicate the high efficiency of the Convolutional Neural Network for solving multiclass problems. Thus, from 1144 samples used for the model testing, only 36 and 44 were identified incorrectly when using the complete and filtered data of gene expression profiles, respectively. The classification accuracy was 97% at using complete data and 96% when using filtered data. Moreover, the analysis of the values of the classification quality criteria by classes shows the high accuracy of identification of samples for which no cancer was detected. This fact confirms the results

of the binary classification obtained using the Random Forest (RF) classification algorithm and indicates that the CNN can be effectively used successfully in systems for identifying the presence or absence of a cancerous tumour based on gene expression data.

Table 7

The results of the application of CNN for the classification of various types of cancer when using filtered data (14451 genes)

Class	Number of samples	Sensitivity	Specificity	F-measure	Accuracy
ACC	24	0.74	0.96	0.84	
GBM	73	0.96	0.92	0.94	
KIRC	169	0.99	0.99	0.99	
LGG	191	0.97	0.98	0.97	96%
LUAD	200	0.94	0.96	0.95	
LUSC	169	0.96	0.93	0.95	
SARC	73	0.97	0.96	0.97	
STAD	96	0.95	0.93	0.94	
NORMAL	149	0.99	0.99	0.99	

Figure 6 shows charts of accuracy and loss function values obtained on a set of objects that were used for training the model validation. The same charts were obtained for filtered data.

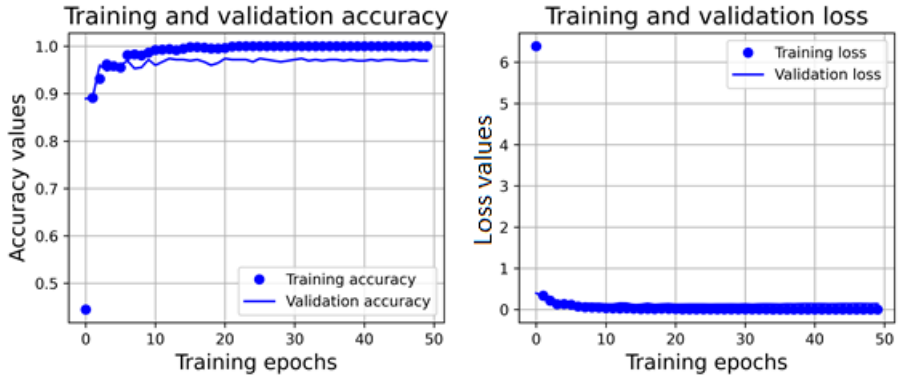

Figure 6: Charts of accuracy and loss function values obtained on the subset of objects that were used directly for network training and model validation using the complete dataset during the model training

When the simulation procedure was implemented, the set of data allocated for training the model was shared into two subsets in the ratio of 0.7/0.3. 70% of samples were used directly for training and 30% for network validation during the training process. Convergence of both the classification accuracy and the loss function obtained on both data subsets indicates the absence of retraining of the network. In this case, the classification results received on the test data can be considered adequate. The analysis of the obtained results also shows a slight discrepancy in the results of the identification of the samples obtained on the complete and filtered subsets. This fact confirms the expediency of removing non-informative genes at the stage of data preprocessing according to statistical and entropy criteria since reducing the certain number of profiles in this case almost does not affect the results of the investigated samples classification'. However, we would like to note that the time for data processing and the amount of computer resources required for data processing are significantly less in comparison with the processing of complete data.

6. Conclusions

In this research, we have presented the stepwise procedure of gene expression data formation, analysis and preprocessing based on the combined use of various techniques of big data processing.

The gene expression data of patients examined for various kinds of cancer from The Cancer Genome Atlas were applied as the experimental ones when the simulation procedure was implemented. This dataset contained nine types of data, eight of which corresponded to eight types of cancer, and the ninth data corresponded to samples for which no cancer was identified as a result of clinical trials. The total number of studied samples was equal to 3269. Each sample of initial data contained, as attributes, 19947 genes.

The procedure of gene expression data formation at the first step assumed the application to the initial data functions of the TCGAbiolinks package in order to transform the values of gene expression in a more suitable range. Then, the non-expressed and lowly-expressed genes for all samples were removed from the data as non-informative ones. At this phase, we removed 682 non-expressed genes, and their quantity was reduced from 19947 to 19265. In the second step, the non-informative genes in terms of various statistical and entropy criteria were removed from the dataset. To identify the genes' significance based on various kinds of criteria, we have applied the Harrington desirability function. Implementation of this phase has allowed us to reduce the genes from 19265 to 14451.

In the final phase, we applied the Random Forest classifier to solve the binary classification task and the Convolutional Neural Network to solve the multi-classification task with calculation classification performance criteria at each phase of this procedure implementation. The analysis of the simulation results indicates the high efficiency of the application of both the RF algorithm and the CNN for solving both binary and multiclass tasks. In all cases, the classification accuracy calculated with the use of the test dataset was very high. From 1144 samples used for the model testing, only 36 and 44 were identified incorrectly when using the complete and filtered data of gene expression profiles, respectively.

The further perspectives of the authors' research are the creation of a hybrid model of gene expression data processing based on the joint use of clustering and classification techniques, where the presented in the manuscript method will use at the stage of data pre-processing.

7. References

[1] H. Salem, G. Attiya, N. El-Fishawy. Classification of human cancer diseases by gene expression profiles. Applied Soft Computing, 2017, vol. 50, pp. 124-134. doi: 10.1016/j.asoc.2016.11.026

[2] F.V. Sharbaf, S. Mosafer, M.H. Moattar. A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. Genomics, 2016, vol. 107(6), pp. 231-238. doi: 10.1016/j.ygeno.2016.05.001

[3] S.A.A. Vijay, P.G. Kumar. Fuzzy expert system based on a novel hybrid stem cell (HSC) algorithm for classification of microarray data. Journal Medical Systems, 2018, vol. 42, art.no. 61. doi: 10.1007/s10916-018-0910-0

[4] N. Almugren, H. Alshamlan. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE Access, 2019, vol. 7, art. no. 8736725, pp. 78533-78548. doi: 10.1109/ACCESS.2019.2922987

[5] H. Lu, J. Chen, K. Yan, et al. A hybrid feature selection algorithm for gene expression data classification. Neurocomputing, 2017, vol. 256, pp. 56-62. doi: 10.1016/j.neucom.2016.07.080

[6] M.A. Khan, M.I.U. Lali, K. Javed, et al. An Optimized Method for Segmentation and Classification of Apple Diseases Based on Strong Correlation and Genetic Algorithm Based Feature Selection. IEEE Access, 2019, vol. 7, art. no. 8675916, pp. 46261-46277. doi: 10.1109/ACCESS.2019.2908040

[7] M. Dashtban, M. Balafar. Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics, 2017, vol. 109(2), pp. 91-107. doi: 10.1016/j.ygeno.2017.01.004

[8] R. Aziz, S.K. Verma, N. Srivastava. A novel approach for dimension reduction of microarray. Computational Biology and Chemistry, 2017, vol. 71, pp. 161-169. doi: 10.1016/j.compbiolchem.2017.10.009

[9] P. Moradi, M. Gholampour. A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. Applied Soft Computing, 2016, vol. 43, pp. 117-130. doi: 10.1016/j.asoc.2016.01.044

[10] I. Jain, V.K. Jain, R. Jain. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Applied Soft Computing, 2018, vol. 62, pp. 203-215. doi: 10.1016/j.asoc.2017.09.038
[11] E. Pashaei, M. Ozen, N. Aydin. Gene selection and classification approach for microarray data based on random forest ranking and BBHA. In Proc. IEEE-EMBS Int. Conf. Biomed. Health Inform. (BHI), 2016, pp. 308-311. doi: 10.1109/BHI.2016.7455896
[12] S.S. Shreem, S. Abdullah, M.Z. Nazri. A Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm. International Journal of Systems Science, 2016, vol. 47(6), pp. 1312-1329. doi: 10.1080/00207721.2014.924600
[13] P. Tumuluru, B. Ravi. GOA-based DBN: Grasshopper optimization algorithm-based deep belief neural networks for cancer classification. International Journal of Applied Engineering Research, 2017, vol. 12(24), pp. 14218-14231.
[14] H. Djellali, S. Guessoum, N. Ghoualmi-Zine, S. Layachi. Fast correlation based filter combined with genetic algorithm and particle swarm on feature selection. In Proc. 5th Int. Conf. Elect. Eng.-Boumerdes (ICEE-B), 2017, pp. 1-6. doi: 10.1109/ICEE-B.2017.8192090
[15] M. Divate, A. Tyagi, D.J. Richard, et al. Deep Learning-Based Pan-Cancer Classification Model Reveals Tissue-of-Origin Specific Gene Expression Signatures. Cancers, 2022, vol. 14(5), art no. 1185. doi: 10.3390/cancers14051185
[16] Y. Xiao, J. Wu, Z. Lin, X. Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. Computer methods and programs in biomedicine, 2018, vol. 153, pp. 1-9. doi: 10.1016/j.cmpb.2017.09.005
[17] M. Mostavi, Y.C. Chiu, Y. Huang, et al. Convolutional neural network models for cancer type prediction based on gene expression. BMC Med Genomics, 2020, vol. 13(5), art no. 44. doi: 10.1186/s12920-020-0677-2
[18] El. Resource. Available on https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
[19] A. Colaprico, T.C. Silva, C. Olsen, et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research, 2016, vol. 44(8), art. no. e71. doi: 10.1093/nar/gkv1507
[20] El. Resource. Available on https://www.bioconductor.org/
[21] R. Ihaka, R. Gentleman. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics, 1996, vol. 5(3), pp, 299-314. doi: 10.2307/1390807
[22] El. Resource. Available on https://www.illumina.com/
[23] I. Liakh, S. Babichev, B. Durnyak, I. Gado. Formation of Subsets of Co-expressed Gene Expression Profiles Based on Joint Use of Fuzzy Inference System, Statistical Criteria and Shannon Entropy. Lecture Notes on Data Engineering and Communications Technologies, 2023, vol. 149, pp. 25-41. doi: 10.1007/978-3-031-16203-9_2
[24] S. Babichev, J. Škvor. Technique of Gene Expression Profiles Extraction Based on the Complex Use of Clustering and Classification Methods. Diagnostics, 2020, vol. 10 (8), art. no. 584. doi: : 10.3390/diagnostics10080584
[25] S. Babichev, J. Krejci, J. Bicanek, V. Lytvynenko, V. Gene expression sequences clustering based on the internal and external clustering quality criteria. Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, 2017, vol. 1, art. no. 8098744, pp. 91-94. doi: 10.1109/STC-CSIT.2017.8098744
[26] L. Yasinska-Damri, S. Babichev, B. Durnyak, T. Goncharenko. Application of Convolutional Neural Network for Gene Expression Data Classification. Lecture Notes on Data Engineering and Communications Technologies, 2023, vol. 149, pp. 3-24. doi: 10.1007/978-3-031-16203-9_1